Molecular approaches for studying the evolution of the Xenacoelomorpha

Helen Elizabeth Robertson

UCL

Submitted for the Degree of Doctor of Philosophy

2017

Declaration

I, Helen Elizabeth Robertson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The Xenacoelomorpha comprises the genus *Xenoturbella* and two related groups of acoelomorph worms – the Acoela and Nemertodermatida. The phylogenetic position of the Xenacoelomorpha is debated. Previous work suggested that they are deuterostomes; alternative publications suggest they are outside the main group of bilaterians (protostomes and deuterostomes). All members of the Xenacoelomorpha have a simple body plan: they are unsegmented; lack a coelomic cavity; and are commonly assumed to lack any of the organs commonly associated with the Bilateria.

In the first part of this thesis, I develop and apply gene visualisation protocols (*in situ* hybridisation and immunohistochemistry) to investigate the expression patterns of genes commonly associated with ultrafiltration in *Xenoturbella* and the acoel *Symsagittifera roscoffensis*. Limited molecular protocols have previously been applied to members of the Acoelomorpha, but none have been successful in *Xenoturbella*. Given their simple morphology, the Xenacoelomorpha have been assumed to lack any structures specialised for filtration or excretion (nephrocytes). The main bilaterian protostome and deuterostome grouping has thus been termed the Nephrozoa, implying that nephrocyte-like structures, required for ultrafiltration, are an innovation of this clade. Understanding the role of these genes in members of the Xenacoelomorpha could shed light on the origin and homology of filtratory structures, which remain unclear. More broadly, establishing reliable gene visualisation approaches in *Xenoturbella* is helpful for better understanding their morphology and organisation.

In the second part, I apply novel RNA-Seq approaches in *Xenoturbella.* These comprise whole-organism single cell sequencing, to investigate organisational complexity and tissue specification in *Xenoturbella*; and 'Tomoseq', for investigating spatially resolved transcriptomics across the anteroposterior body axis. In association with my gene visualisation

protocols, these techniques are valuable approaches to enhance our understanding of the biology and complexity of *Xenoturbella*.

Statement of Scientific Impact

The breadth of the animal kingdom has been a source of fascination for hundreds of years. Where did the myriad of forms that we see across the world originate from, and how did such diversity arise? For evolutionary biologists, the species belonging to the Xenacoelomorpha represent a particular conundrum in our current understanding of the evolution of the Bilateria. Where these enigmatic marine organisms belong in comparison to their bilaterian relatives, and how their cryptically simple morphology can be interpreted, are fascinating questions that are not easy to answer.

Investigating the cellular complexity of members of the Xenacoelomorpha, and the expression of genes that are known to be well-conserved in other organisms, is valuable to the study of evolutionary biology for a number of reasons. Not only will it better-inform our understanding of the organisation of these animals, but in widening our investigation away from typical model organisms, we can begin to bring together the history of when different genes acquired new functions and when different cell types and tissues emerged. For evolutionary developmental biology, making these comparisons on a broad scale is fundamental – and the Xenacoelomorpha are a particularly interesting group to target for investigation. Ultimately, this basic research contributes another piece in the puzzle of how the diversity of life on earth arose.

The protocols I developed and used in this thesis to understand more about the cryptic Xenacoelomorpha are novel and exciting new techniques in molecular biology. Single cell sequencing and differential transcriptomics have applications far beyond the scope of evolutionary biology: for example, single cell sequencing is already being used to investigate disease development and progression, and response to drug treatment. By applying this technique to animals for which molecular protocols are not well established, we can fine-tune and optimise their implementation – with wider benefits for their future application across the scope of biology and medical

sciences. Basic research is at the core of advancing fundamental knowledge, and the surprising nature of research often means that these investigations ultimately shape the future of applied science. Using research to advance knowledge, be that in the context of evolutionary biology, or across the breadth of biological and medical sciences, should perhaps itself be the most important reason for it to be undertaken.

Acknowledgements

Firstly, I would like to thank my supervisor Max Telford for all the help, opportunities and advice that he has given me in the time since I started my undergraduate degree eight years ago. My interest in evolutionary and molecular biology has undoubtedly been shaped by his work and guidance, and for that I am very grateful. I would also like to thank him for the opportunities that this PhD has afforded me – both in terms of travel and the innovative new approaches that I have been able to use during the course of my research. Many thanks also to Paola Oliveri for her scientific guidance, discussion, and help with molecular protocols.

I would like to thank members of the Telford and Oliveri labs from 2013-2017: Marta, Johannes, Anne, Philipp, Steven, Fraser, Irepan, Anna, Natalie and Laura. Your advice, and friendship in and out of the lab has made my PhD a brilliant experience. In particular, thanks to Anne for making our field-work in Sweden one of the highlights of my PhD; and to Philipp for his scientific discussion, and incredible patience and help with bioinformatic analysis. I have also had the opportunity to work with a number of other institutions and labs over the course of my PhD, and their knowledge and specialisation have been of huge benefit to my research. Thanks to staff at the Sven Lovén Centre in Sweden for their assistance on collection trips; to Pawan Dhami of the UCL Cancer Institute for her help with sequencing; and Heather Marlow and lab members at the Pasteur Institute for our collaboration on the single cell sequencing work.

I am fortunate to have had brilliant support from my friends and family throughout my PhD, for which I am immensely grateful. Thank you to my parents and brother for their constant encouragement - and for believing in me more than I sometimes do myself. And to Nat, Jo and Kathryn, for your wonderful friendship and always enquiring about 'the worms'.

Finally, I would like to thank Alex for his unwavering support and reassurance throughout my PhD and in the final months of thesis writing. Your patience and support means the world to me and I owe you very much. Thank you.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Reconstructing the tree of life: morphological vs. molecular data and the new animal phylogeny

1.1.1 General Introduction

The animal kingdom comprises some 1.3 million described species, made up of a hugely diverse assortment of forms. Ever since Darwin and Haeckel described their original ideas of relating these species in a branching 'tree of life', the Metazoa and the variety it represents has endlessly fascinated the scientific community. In the time since the publication of *Origin of Species* and *Generelle Morphologie der Organismen*[1], scientists have endeavoured to understand the evolutionary history of metazoan species. Establishing the interrelationships between the branches of the tree remains a challenge for both morphological and molecular biology.

1.1.2 Morphological data

The first century of phylogenetics was defined by a morphology-based approach. How could the relationships between organisms be unravelled based on the study of shared body plans, morphological features, and early developmental characteristics? Interpreting how different morphologies are related was at the centre of this approach, and much debate surrounded the problem of identifying homologous structures amongst different organisms, and inferring how they had evolved or differentiated between different groups. Although there was often no clear consensus for the interpretation of homologous structures, two general themes pervaded morphology-based investigations. First was the seemingly logical idea that animal body plan progressed from that of simple to more complex during evolution – that an increased number of germ layers, a change from radial to bilateral body symmetry, and an increased specialisation of cell types and organ systems

was indicative of increased evolutionary complexity. Second was a focus on understanding the features that were likely present in the common ancestor of the bilaterally symmetrical animals: which characters observed in extant Bilateria were primitive, and which were derived[2]. Unsurprisingly, the differing interpretations of primitive vs. derived features had a significant impact on the topology of the tree. Placing different extant body plans and modes of development at the base of the tree implied that other phyla had derived from these early character states, and there was no clear consensus as to which taxa could confidently be assigned as representing a primitive or ancestral morphology[3].

Nonetheless, some of the main branches of phylogenies remain credible in modern phylogenetics. Indeed, the grouping of different taxa during the late 19th Century into their respective phyla – a classification determined by a set of characters unique to that group –are still largely held to be true today. However, understanding how different phyla are related to one another cannot easily be achieved based solely on morphological and developmental features. Whilst some of the early morphology-based predictions of interrelatedness still hold true today, others have been re-evaluated by the focus on molecular data in modern evolutionary biology[4].

1.1.3 Molecular data

Molecular tools and technologies have had a significant impact on our understanding of metazoan relatedness. The first efforts to reconstruct phylogeny using genomic data were based on the nuclear small subunit (SSU) gene – but were confounded by limited availability of data, and systematic errors in tree reconstruction[5]. Despite this, these early efforts marked a turning point in our understanding of animal phylogeny. Over the past 30 years, a succession of 'genomic revolutions' has re-shaped our understanding of the interrelatedness of metazoan taxa.

Molecular data – including nucleotide or amino acid sequences, or microRNA data - offers a number of advantages for phylogenetic

reconstruction compared to morphological observations. Genes are far more numerous than phenotypic characters, and each gene itself comprises up to thousands of nucleotides or amino acids, offering a far-expanded data set for comparison[3]. This also decreases the likelihood of misidentifying convergent evolution as primary homology: the probability of thousands of nucleotides being identical by chance is very low, whilst examples of convergent phenotypes are numerous across the Bilateria, and are not indicative of evolutionary closeness. Lastly, comparing molecular data is theoretically more straightforward than comparing morphological characters: whilst larger data sets are available, these are constrained to limited character states (four different nucleotides; 20 amino acids; or 64 codons)[2]. This is not to say that molecular data are without their limitations, and molecular phylogenies are prone to both systematic and stochastic errors that can have significant implications for the misplacement of phyla.

1.1.4 The new animal phylogeny

Our understanding of molecular phylogeny, and the knowledge contributed by evolutionary developmental (evo-devo) studies, means that much of the metazoan tree is well resolved, with consistent support from different sampling methods, taxon choice, and phylogenetic reconstruction methods. Monophyly of the animal kingdom is well supported, and all extant animals belong to one of five monophyletic clades: Ctenophora (comb jellies); Cnidaria (corals and medusa); Porifera (sponges); Placozoa; or Bilateria (all other bilaterally symmetrical animals)[6].

The Bilateria comprises the most diverse of all metazoan lineages – both in number and form – and understanding the relationships between bilaterian phyla has been the focus of countless molecular phylogenies. The so-called 'new-animal phylogeny' supports the monophyly of the Bilateria[7]. The presence of two main bilaterian clades, the Protostomia and Deuterostomia, is also well supported (Figure 1.1).

For the most part, there is a general consensus for the composition and interrelatedness of members of Deuterostomia. The deuterostomes are considered to include two main groupings: the Ambulacraria (Hemichordates and Echinoderms) and the Chordata (Vertebrata, Urochordata and Cephalochordata).

Other bilaterian members are assigned to the Protostomia, which itself comprises the monophyletic groupings Ecdysozoa and the Lophotrochozoa. The Ecdysozoa – including the arthropods, nematodes, and priapulids – are united based on the common presence of a moulting cuticle (ecdysis); the Lophotrochozoa, including the annelids, molluscs, and a number of other protostome phyla, are so-called owing to the commonly shared characteristics of either a ciliated feeding structure (lophophore) or a trochophore-type larvae. The split of protostome members into ecdysozoans and lophotrochozoans is a widely-cited example of molecular data overturning ideas of phylogeny originally founded on morphology: the grouping of animals as different as the arthropods and the pseudocoelomate nematodes is not easily determined by their diverse morphologies. Prior to this conclusion, the nematodes were typically placed as a primitive, early branch within the Bilateria, and the arthropods placed more closely to the vertebrates in the Coelomata, and more specifically with the annelids in the Articulata. Certainly the Ecdysozoa represent a group where exclusively morphological observations can be misleading: numerous lines of molecular evidence support Ecdysozoa over the Articulata or Coelomata, but this is not clear when we consider the presence or absence of a complex character, such as segmentation.

Our re-evaluation of protostome phylogenetics emphasises the shortcomings of considering 'simple' characters – such as the absence of a coelom – as being indicative of evolutionary primitiveness. Phylogenies derived from molecular data place seemingly 'simple' taxa - including Platyhelminthes, Nemertea and Nematoda – amongst more 'complex' coelomate taxa, whereas many morphological analyses placed these acoelomate or pseudocoelomate groups together near the base of the

Bilateria. Reconsidering simplicity not only has implications for the position of taxa within the Bilateria, but also suggests that simplification could be a progressive driving force in evolution.

Despite significant progress in recent years in our understanding and interpretation of phylogeny, molecular data have not been conclusive in resolving all of the nodes of the bilaterian tree. Of the remaining questions in bilaterian evolution, the enigmatic members of the so-called Xenacoelomorpha represent perhaps one of the most intriguing. These unassuming worms – comprising the genus *Xenoturbella* and the two related groups of acoelomorph worms Acoela and Nemertodermatida – have been the focus of countless molecular analyses in recent years, and pose a fascinating conundrum in our understanding of bilaterian evolution.

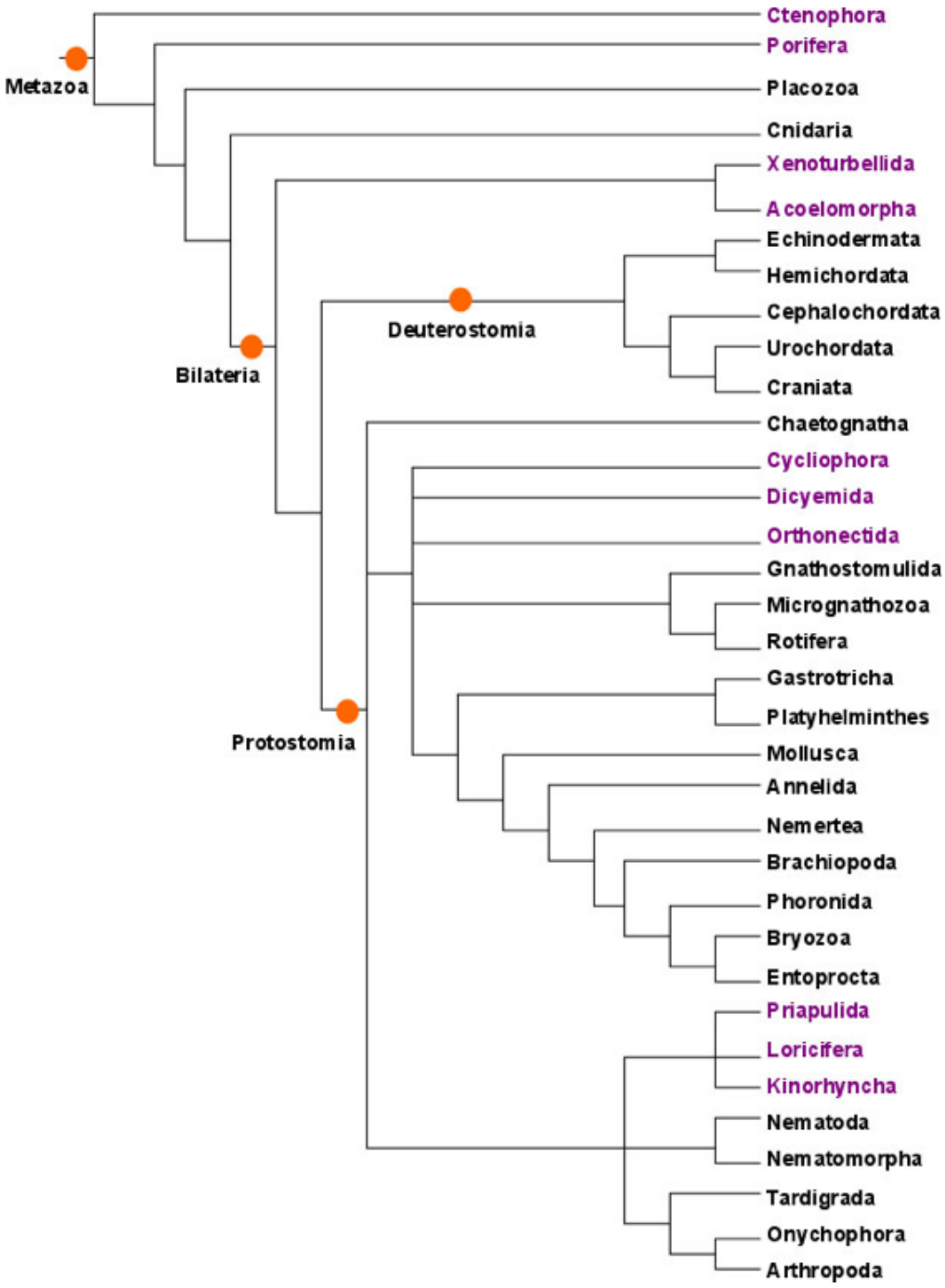


Figure 1.1. The 'new animal phylogeny'. Figure adapted from Giribet (2015)[8]. Taxa in purple indicate those for which phylogenetic position remains uncertain or poorly supported. Major clades in the Metazoa indicated by orange circles.

1.2 Introduction to the Xenacoelomorpha

1.2.1 Acoelomorpha

Acoel flatworms (phylum Xenacoelomorpha/Acoelomorpha, class Acoela) are small (~1mm), soft-bodied, unsegmented worms lacking a gut epithelium, coelomic cavity and anus. Instead, they possess a ventral mouth opening and a simple syncytial digestive system, with the space between their primitive gut and body wall filled with parenchymal cells and occasionally vacuoles and gland cells[9]. As is found for many microscopic worms, acoels glide via a ciliated epidermis[10]. A unique morphological feature of the Acoela is the presence of a balance and sensory receptor called the statocyst: a sac-like structure that encompasses a mineralised mass (statolith) surrounded by sensory hairs called setae, which allows the animal to sense change in orientation and maintain balance[10] (Figure 1.2).

Acoels are a diverse taxon, comprising greater than 400 different species, described from the littoral and sub-littoral zones of predominantly marine ecosystems across the globe. Most species are free-living, although seven described species are parasites or endosymbionts found living in the digestive system of echinoderms. Their pigmentation and body shape is diverse, and varies according to their habitat – species that live in coarse sand are typically long and slender, whilst those living in the mud are more compact and tear-drop shaped[10] (Figure 1.3).

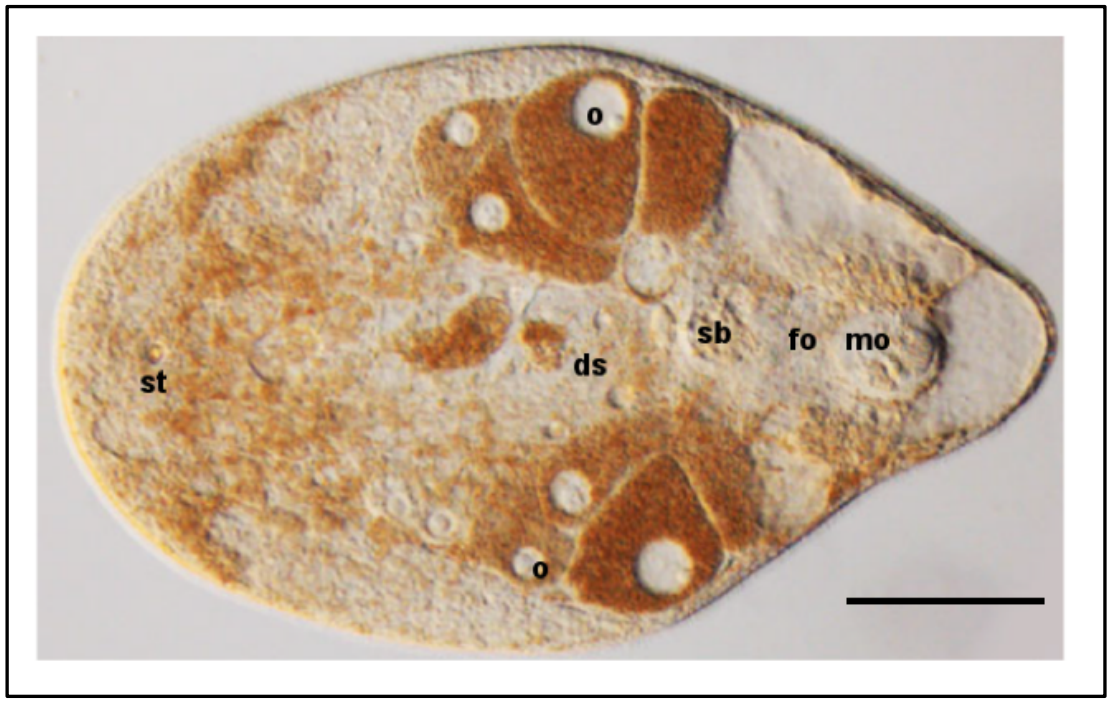


Figure 1.2. General morphology of the Acoela as shown by *Isodiametra pulchra*. Figure adapted from Perea-Atienza *et al.* (2013)[11]. st = statocyst; ds = digestive syncytium; o = oocyte; sb = seminal bursa; fo = female copulatory organ; mo = male copulatory organ. Scale bar: 80μm.

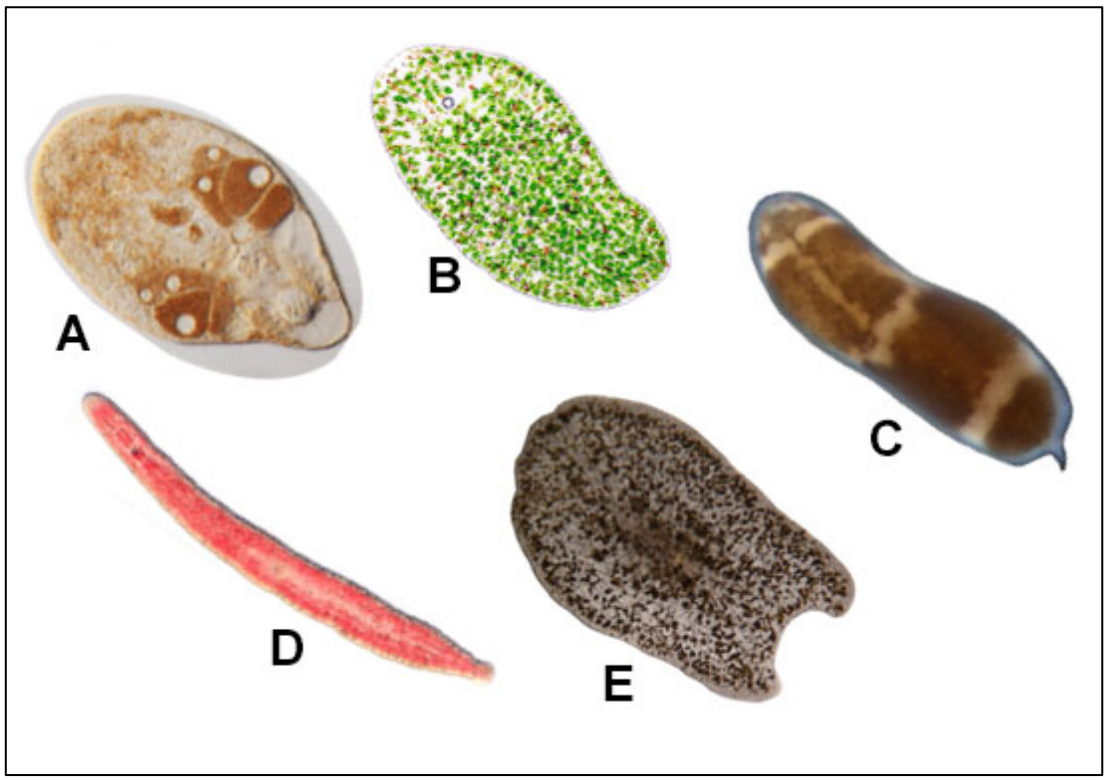


Figure 1.3. The diversity of the Acoela. (A) *Isodiametra pulchra* (image Perea-Atienza *et al.* (2013)[11]); (B) *Symsagittifera roscoffensis* (image Arthur Hauck, https://www.rzuser.uni-heidelberg.de/~bu6/Convoluta.html); (C) *Hofstenia miamia* (image Mansi Srivastava http://www.srivastavalab.org/research.html); (D) *Paratomella rubra* (image Bernhard Egger); (E) *Waminoa sp.* (image Mijgerde *et al.* (2012)[12]). Anterior to the left in all images.

Due primarily to their common acoelomate body, and the shared absence of a through gut, Acoela were traditionally grouped alongside the Nemertodermatida within the class Acoelomorpha, as an order within the Platyhelminthes. The first molecular systematic studies on these animals using small subunit (SSU) ribosomal RNA gene sequences in fact demonstrated that the Acoelomorpha were a lineage quite separate from the main clade of the Platyhelminthes (Rhabditophora and Catenulida)[13-15]. Instead, these initial molecular studies supported a position of Acoelomorpha as diverging from the rest of the Bilateria prior to the protostome/deuterostome common ancestor. This early-branching position of the Acoelomorpha has subsequently been supported by consideration of a number of molecular characters including *Hox* gene signatures[16,17]; combined SSU+LSU studies[18]; and mitochondrial genome data[19,20], amongst others. If this phylogenetic position of the Acoelomorpha as the most basally branching, triploblastic bilaterian is correct, it would place them in a pivotal position between diploblasts (Porifera, Cnidaria and Ctenophora) and the main bilaterian clade (Protostomia and Deuterostomia), implying a critical role for them for our understanding of the evolutionary history of the Metazoa. Acoels have, in consequence, been studied by evolutionary and developmental biologists to provide an insight into the morphology, genetics and development of the most recent common ancestor of the Bilateria[21].

1.2.2 *Xenoturbella*

Along with the Acoelomorpha, the equally simple but somewhat more enigmatic marine worms in the genus *Xenoturbella* pose an ongoing phylogenetic conundrum. *Xenoturbella bocki* is a small (~2cm long), yellowy-orange flattened worm with darker brown pigmented spots across its body. *Xenoturbella* was originally described in 1949 from specimens collected in mud off the west coast of Sweden[22]. From a morphological perspective, *Xenoturbella bocki* appears to be incredibly simple. It moves via gliding on a ciliated epidermis; has a basiepithelial nerve net instead of a complex nervous system; has no coelomic cavity; and is assumed to lack any other defined organ systems (Figure 1.4). The most obvious morphological features of *Xenoturbella* are its anterior statocyst (much like that found in the acoels); a radial furrow around the mid-section of the animal; and more cryptic anterior lateral furrows that have been hypothesised to have a sensory function. Apart from these characters, very little is known regarding the degree of organisation of cell types or putative tissue types in *Xenoturbella*.

For many decades, *X. bocki* was thought to be the lone representative of the genus *Xenoturbella*, and so the description in 2016 of four new *Xenoturbella* species from the East Pacific, living on the sea-bed at depths of between 600 and 3700 metres, was a pivotal finding[23]. These colourful new species are comparatively large - ranging to greater than 20cm in length – but their simple body plan and external appearance make them clearly recognisable as morphologically similar to *X. bocki*. Furthermore, mitochondrial genome data confirm their placement in the same genus, with an apparent split between 'deep' species, living at depths of between 700 and 3700 metres, and 'shallow' species – including *X. bocki* – living at less than 650 metres (Figure 1.5). This molecular analysis also found very little difference between the five species, indicating that they had diverged comparatively recently[23].

Figure 1.4: *Xenoturbella bocki* collected from the Gullmarsfjorden, west coast of Sweden. Ventral furrow indicated by arrowheads. Anterior at the top in all panels. Left panel: lateral view; middle and right panel: ventral view.

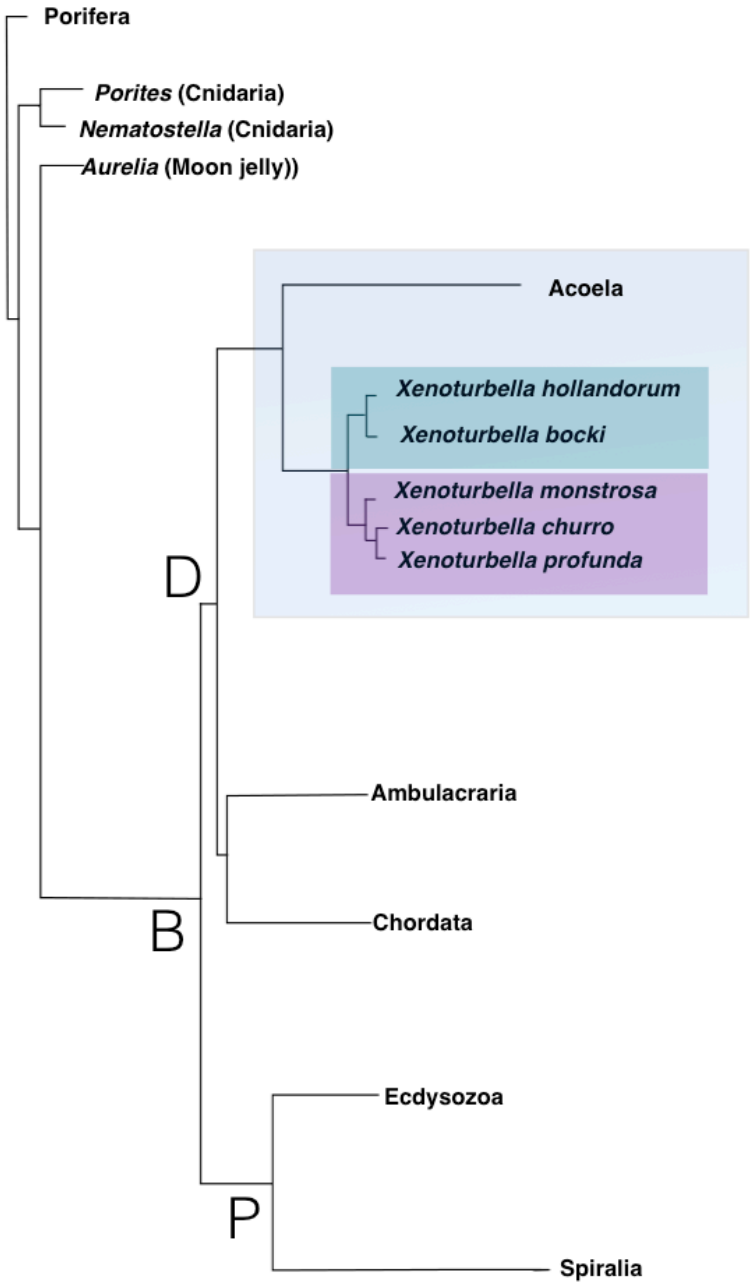


Figure 1.5. Interrelatedness of *Xenoturbella.* Phylogeny based on mitochondrial protein-coding genes, adapted from Rouse *et al.* (2016)[23]. Xenacoelomorpha (*Xenoturbella*+Acoela) shown by blue box. *Xenoturbella* species found to split into two groups of 'shallow' species (green) and 'deep' species (purple). B = Bilateria; D = Deuterostomia; P = Protostomia.

Much like the Acoelomorpha, the simple structure of *Xenoturbella* has, historically, proven to be a potentially misleading evolutionary clue. The first genetic data from *X. bocki* surprisingly linked them to the Mollusca[24]. This was an unexpected result considering the disparity between the body plans of *Xenoturbella* and molluscs, and was subsequently found to be the result of an artefact, derived from the bivalve molluscs – on which it preys - remaining in the gut of *X. bocki*[25]. Subsequent multi-gene data sets using molecular data confidently derived from *X. bocki* instead grouped them alongside the Ambulacraria within the Deuterostomia. Later, molecular analyses linked them to the Acoelomorpha: a relationship that makes sense of their shared simple body plan and other specific morphological characters including ciliary ultrastructure[26] and nervous system[27]. In addition, *Xenoturbella* share with the acoels the feature of a balancing and navigational statocyst. That the Acoelomorpha and *Xenoturbella* can be grouped together in the Xenacoelomorpha is where our confident assessment of their evolutionary history ends.

1.2.3 Phylogenetic position of the Xenacoelomorpha

Two main hypotheses pervade regarding the placement of Xenacoelomorpha within the Metazoa: either as 'basal' bilaterians, separate from the main protostome and deuterostome grouping (which has been termed the 'Nephrozoa'), or that Xenacoelomorpha are in fact deuterostomes, forming the sister phylum to the Ambulacraria (comprising Echinodermata and Hemichordata) (Figure 1.6). The affinity to the deuterostomes was primarily based on three lines of evidence, comprising mitochondrial genome sequences; microRNA complements; and nuclear genes[28]. Recent phylogenetic analysis of mitochondrial genes from the four newly reported *Xenoturbella* species also find a deuterostome affinity, but with relatively low support[23]. The placement of Xenacoelomorpha as a basal bilaterian has been supported by ESTs and a number of analyses of transcriptomic data[21,29]. The most recent analysis of RNA-Seq data from seven acoel species, four nemertodermatids (together forming the

Acoelomorpha), *Xenoturbella bocki*, and >60 other diverse metazoan taxa represents perhaps the most comprehensive data set in terms of Xenacoelomorpha phylogenetics, and consistently found support for a basal position, separate from the Nephrozoa[30]. Nonetheless, systematic errors have pervaded phylogenetic reconstruction involving Xenacoelomorpha, and increased taxon sampling alone is perhaps not sufficient to overcome these problems and reach a definitive conclusion.

In either position, the Xenacoelomorpha represent an interesting phylum for our interpretation of animal evolution (Figure 1.6). They are recognised to have a very simple morphology, and their simple body plans mean that they are assumed to lack many of the tissue types or cellular differentiation otherwise associated with bilaterian animals. All multicellular animals comprise cell types specialised for discrete functions: non-bilaterians including the cnidarians, ctenophores and poriferans have historically been considered to have relatively few cell types, and cell type number has frequently been used as a marker for organismal complexity. Consequently, placing xenacoelomorphs at the base of the Bilateria makes sense of their simple morphology and the assumed absence of numerous cell or tissue types. This position would indicate that they branched from the main protostome and deuterostome group of the bilaterians early in evolution, and imply that they diverged prior to the innovation of more 'complex' structures or cell types that we typically associate with bilaterians.

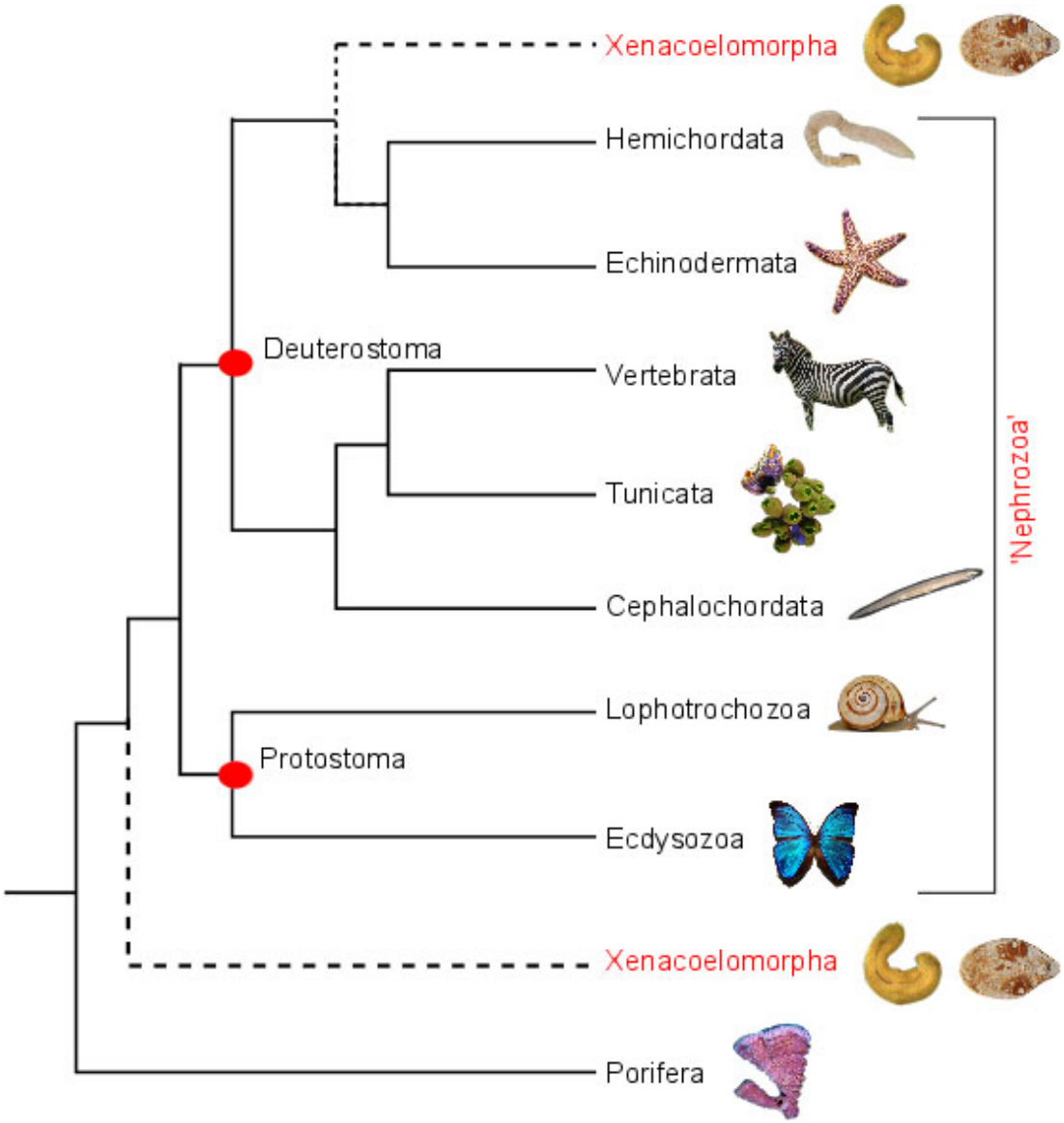


Figure 1.6. The placement of Xenacoelomorpha in the Bilateria. Two alternate theories pervade regarding the placement of the Xenacoelomorpha, as indicated by dashed lines: (1) as sister group to the Ambulacraria within the deuterostomes; (2) as basal bilaterians, separate from the main protostome and deuterostome grouping.

However, morphological simplicity can arise through loss of characters, and their simple body plan may not be evidence for a basal position. Furthermore, our limited knowledge of their morphology and development means we cannot infer a great deal about the cell or tissue types they possess. In an alternative evolutionary scenario, Xenacoelomorpha would have descended from the same complex common ancestor of Ambulacraria and Chordata – implying a significant amount of secondary simplification as opposed to primary absence. Both placements are intriguing: either the Xenacoelomorpha could provide an insight into the morphology and complexity of Urbilateria, or studying this phylum could inform our understanding of a fascinating example of the loss of complexity.

1.3 The 'Nephrozoa' and the evolution of excretory systems

1.3.1 The Nephrozoa

Of all the complex structures thought to be missing in the Xenacoelomorpha, the absence of nephridia – that is, an ultrafiltratory and excretory system – has been used to lend evidence to their basal bilaterian position. Owing to the assumed absence of ultrafiltratory structures in the Xenacoelomorpha, all other bilaterians (that is, the protostomes and deuterostomes) have been termed the Nephrozoa. In the following section I will outline the form and function of filtratory systems across the Bilateria.

1.3.2 Homeostasis and excretion

A common requirement for all metazoan organisms is the need to excrete metabolic waste: water, carbon dioxide, and the toxic nitrogenous end products of metabolism, of which ammonia is the primary component. Closely associated with excretion is the requirement for osmoregulation – the maintenance of a constant body fluid volume despite fluctuating ion concentrations, and the regulation of an optimal solute concentration for metabolism during changing body fluid levels. As Metazoa evolved in an aquatic environment, osmoregulation was – and is – of primary importance

for maintaining cellular integrity, metabolic processes, and ionic and acid/base balance.

1.3.3 Excretion in the diploblasts

In diploblastic animals (Porifera, Placozoa, Cnidaria and Ctenophora), organs specialised for excretion are unnecessary. All exchanges with the aquatic environment – including those necessary for osmoregulation – occur directly at the cellular level via simple diffusion from cells directly into sea water much as their single-celled ancestors must have done. Nitrogenous waste can diffuse out of cells directly as ammonia, negating the requirement for any modification of unwanted metabolites prior to their excretion (Figure 1.7).

The loose, cellular-level organisation of Porifera – with a very low degree of closure to environment – means that waste products can be excreted directly. The water canal system of adult sponges, comprising exopinacocytes, endopinecocytes and choanocytes, creates a large surface area of cells that is exposed to the water flowing through them, facilitating osmoregulation and excretion.

In Cnidaria, tissue layers are largely exposed to the environment, with the exception of a small proportion of cells enclosed within the mesoglea between the outer epidermis and the inner gastrodermis. Nonetheless, direct contact between the cells of the tissue layers and the surrounding water, and the short diffusion distance between the two, means that interaction with the external environment can also occur at the cellular level. Likewise, the epithelial organisation of ctenophores allows for exchange with sea water to occur over the surface area of the animal via simple diffusion, meaning that specialised structures for waste excretion are not required (Figure 1.7).

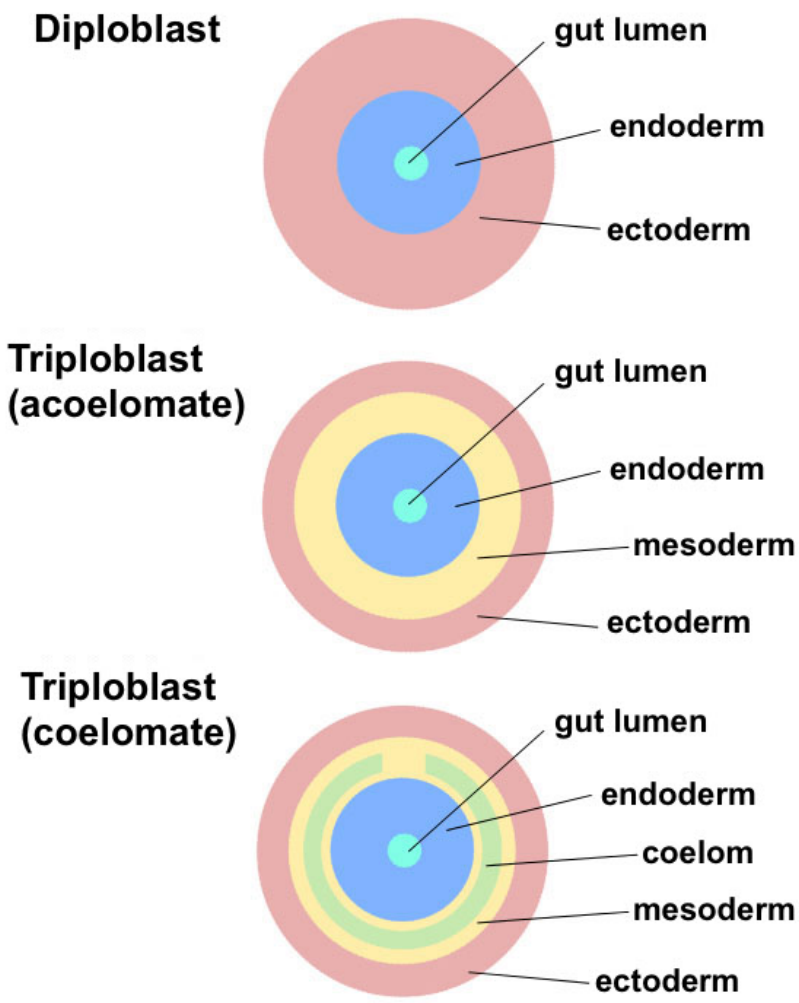


Figure 1.7. Increase in body organisation complexity. Diploblastic organisms (including Cnidaria, Ctenophora and Porifera) are defined by the presence of two germ layers: the endoderm and ectoderm. Bilaterian organisms are triploblastic, comprising three germ layers by the end of gastrulation: the endoderm, ectoderm and mesoderm. A fluid-filled cavity within the mesoderm, the coelom, is found in most triploblastic organisms. Platyhelminthes (the flatworms) and Nemertea (ribbon worms) lack a coelomic cavity and are described as acoelomate. Some other invertebrates, including members of the Nematoda, have a body cavity that is only partly surrounded by mesoderm and are described as pseudocoelomate. The Mollusca, Arthropoda, Annelida, Echinodermata, Chordata and Hemichordata all have a coelom.

1.3.4 Consequences of the evolution of the mesoderm

In the lineage leading to the Bilateria, a consequence of the evolution of a compact, triploblastic organisation – including epithelia and often also blood systems and a body cavity - is the removal of direct contact between internal cells and the external environment[31]. In triploblasts, the mesodermal layer is, in effect, a self-contained compartment, delimited by the outer epidermis and inner digestive tract (Figure 1.7). As a result, the majority of cells in this germ layer never come into contact with the external environment[32]. This loss of contact with the outside world has significant implications for the maintenance of a constant internal environment – that is, homeostasis – and necessitated the evolution of specialised cells and organs for the formation, transport and excretion of metabolic waste, and for osmoregulation.

Of all the end products of metabolism, the toxicity of nitrogenous waste makes it particularly necessary – and challenging – to excrete. Accumulation of ammonia (gaseous, NH_3 and ionic NH_4^+) is particularly harmful, and it must therefore be excreted or incorporated into secondary compounds in order to maintain a tolerable concentration within the organism. The filtration and excretion of ammonia is a hugely costly endeavour in terms of an organism's water budget – its high solubility means that diluting a gram of ammonia to non-toxic concentrations requires 400ml of water[33].

For aquatic organisms (most species of most phyla), water availability is not a limiting factor, and ammonia can be excreted without modification. However, for terrestrial organisms, ammonia must be metabolised into secondary, less toxic compounds such as urea and uric acid (which requires 50 times less water for excretion), which allow them to manage nitrogenous waste excretion without a detrimental impact on their water status[33].

Consequently, coupled with the requirement for excretion in triploblastic organisms is the need for water and solute homeostasis. Both of these challenges – maintaining osmoregulation, and filtering and eliminating toxic

nitrogenous waste products – are managed in bilaterians within the same structures, comprising the excretory system (Figure 1.8). The excretory organs, collectively referred to here as nephridial systems or nephridia, are believed to be a bilaterian innovation that must have been fundamental to the evolution of the mesoderm.

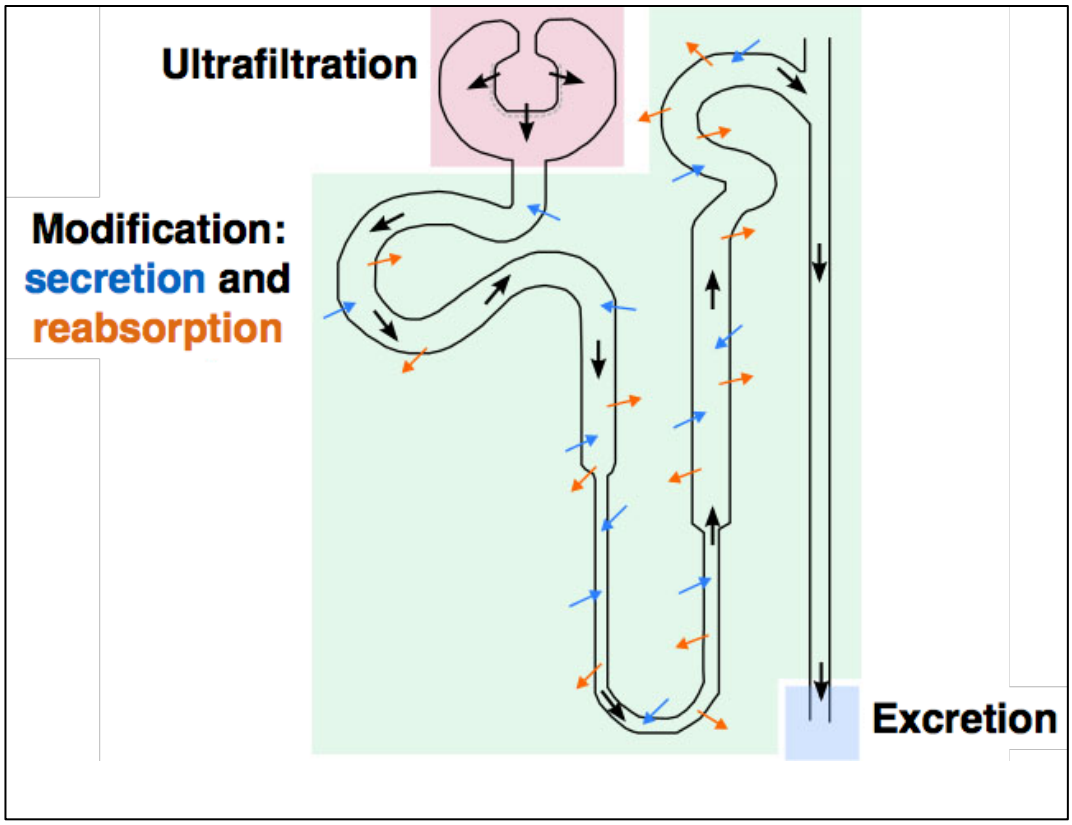


Figure 1.8. Overview of the main components of the generalised bilaterian nephridial system. Schematic based on the vertebrate nephron. This comprises regions of: (1) Ultrafiltration (pink), mediated by either the beating of cilia or pressure exerted by body musculature; (2) Modification (green), facilitated by active transport of solutes into (secretion, blue arrows), or out of (reabsorption, orange arrows) the initial filtrate; (3) Excretion (blue) of the modified filtrate to the exterior of the organism.

1.3.5 Investigating functional conservation of nephridial systems

Despite a broad diversity in the structure of excretory systems that are present in the Bilateria, their functions are largely consistent – the excretion of metabolic waste, and osmoregulation (Figure 1.8) – carried out via two underlying processes:

1) the passive ultrafiltration of blood or body fluid across a 'molecular sieve', and
2) the active modification of this initial filtrate, prior to its excretion from the body.

Of the nephridial systems present across the Bilateria, it is unsurprising that two of the best-studied are those found in vertebrates and in *D. melanogaster*. In addition, the excretory system in the flatworm *S. mediterranea* is well characterised and has helped to inform us about pathways involved in the development, regeneration and function of excretory systems[34]. Conveniently, these well described excretory systems include representatives from all major bilaterian groups: the Deuterostomia (various vertebrate models), Ecdysozoa (*Drosophila melanogaster*) and Lophotrochozoa (*Schmidtea mediterranea*), making them valuable for comparison and to enable us to infer the characteristics of the ancestral bilaterian excretory system.

Whilst all carry out the functions of ultrafiltration followed by modification, at first glance, the nephridial systems present in these taxa – and more widely across the Bilateria – show a myriad of structural variation. The vertebrate kidney and the garland cells and Malpighian tubules of flies, for example, seem irreconcilably different. It is perhaps more useful, therefore, to define a nephridium by their common functional elements, rather than hoping to find a common morphological structure.

The primary filter in nephridia is the extracellular matrix (ECM), which functions as a highly selective, but unspecific, 'molecular sieve'. Body fluid is

driven across the ECM either by a cilia-mediated current/pressure difference, or by pressure exerted by the body musculature. Solutes are excluded by the ECM on the basis of their size and/or electrostatic charge. The ECM retains larger particles, such as large proteins, whilst smaller particles including proteins <~10kDa and nitrogenous waste are able to pass across the filter. The formation of the primary filtrate is 'unselective' in that solutes passing across the ECM are filtered solely based on size or charge, and not by molecular type. In the vertebrates and *D. melanogaster*, specialised epithelial cells (podocytes and nephrocytes, respectively), carry out ultrafiltration. In *S. mediterranea*, filtration occurs in the 'flame cells', which are typical of the Platyhelminthes. These cylindrical cells form a fenestrated barrel around a bundle of cilia, which beat like a 'flame' to draw fluid across the ECM, thus giving rise to their name.

The primary solvent produced by this unspecific filtration is subsequently modified by the reabsorption (endocytosis) and secretion (exocytosis) of solutes in a tubule element, via ATP-dependent active transport. This modification stage is also integral to the function of osmoregulation. Specialised transmembrane proteins called solute carriers (SLCs) carry out a large proportion of solute transportation during this modification stage[35]. SLCs comprise 43 different families and >300 genes, and are categorised into two groups: transmembrane transporters that function by facilitative diffusion; and secondary active transporters, which allow solutes to travel against their electrochemical gradients via coupling to a secondary solute[35]. 'Non-SLC' transporters involved in solute modification include ATP-ases and aquaporins, which are the water channels responsible for osmoregulation[36]. Comparisons between the excretory systems found in the vertebrate and in the flatworm *S. mediterranea* suggest that the distribution of SLCs and aquaporins appears to be commonly organised into different domains within the tubule elements[37]. For example, proximal tubule regions are commonly specialised for recovering solutes that are removed during ultrafiltration: accordingly, SLC expression is highest and most diverse in this region. Distal tubule regions are commonly responsible for acid-base balance and are enriched for SLCs related to bicarbonate and ammonium

transport (*slc4* and *slc42* respectively)[35,37]. The vertebrate distal tubule is also characterised by high aquaporin expression to regulate water absorption[38].

1.3.6 Types of excretory system

From a morphological perspective, nephridial systems have been widely classed into two types across the Bilateria: protonephridia and metanephridia (Figure 1.9). In addition to these two principal types (described in detail below), several groups have seemingly unique excretory systems, and some marine taxa appear to lack a defined excretory system.

1.3.6.1 Protonephridia

In a protonephridium, the proximal end of the nephridial tube is 'blunt-ended', capped by a cup-shaped terminal structure which functions as the site of ultrafiltration (Figure 1.9). In all protonephridial systems, this blunt-ended terminal structure lies in the connective tissue compartment: either the pseudocoel, the blastocoel or, in acoelomates, within the interstitium. The distal portion of the terminal structure acts as a support for the ECM, and is punctuated by pores of ~35-40nm in diameter[37,39]. At these pores, the ECM is the only barrier between the intercellular body fluid and the inside of the nephridial lumen. Solutes within the body fluid are filtered according to the size of the molecules that are able to pass between the pores in the terminal structure and across the ECM into the terminal lumen. In typical protonephridia, the beating of cilia or a flagellum inside the terminal structure induces negative pressure, which draws intercellular fluid across the ECM[40] (Figure 1.10). Following ultrafiltration, the primary filtrate passes directly from the terminal structure into an adjoining tubule, where it is modified by duct cells prior to its excretion via the nephridiopore at the distal end of the protonephridium. As exemplified in *S. mediterranea*, protonephridia have no spatial separation between the site of filtration and the site of modification[37] (Figure 1.10).

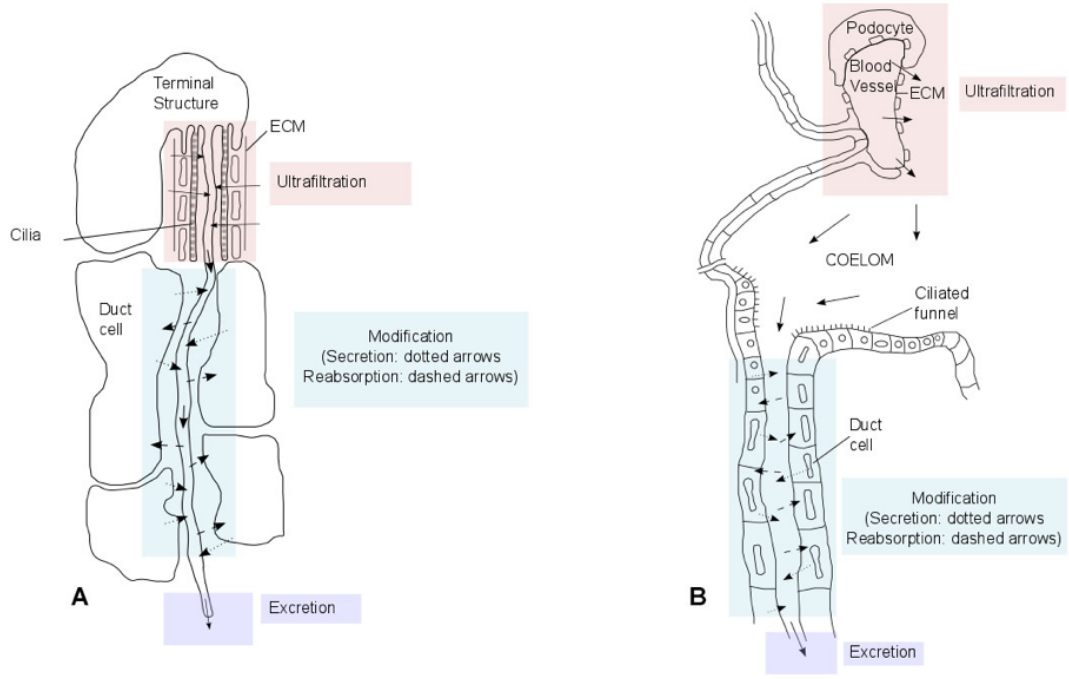


Figure 1.9. Generalised protonephridium and metanephridium. Figure adapted from Bartolomaeus & Ax (1992)[41]. Regions of ultrafiltration (pink); modification (blue) and excretion (purple) shown in both. (A) Protonephridia: ultrafiltration occurs in the blunt-ended terminal structure; filtrate passes directly into the lumen region of an externally opening tubule, where solutes are secreted or reabsorbed by the duct cells to modify its composition prior to excretion. (B) Metanephridia: ultrafiltration occurs in specialised epithelial cells called podocytes (or nephrocytes); ultrafiltrate passes into the coelomic cavity where it is conveyed into the metanephridium via cilia and modified prior to excretion.

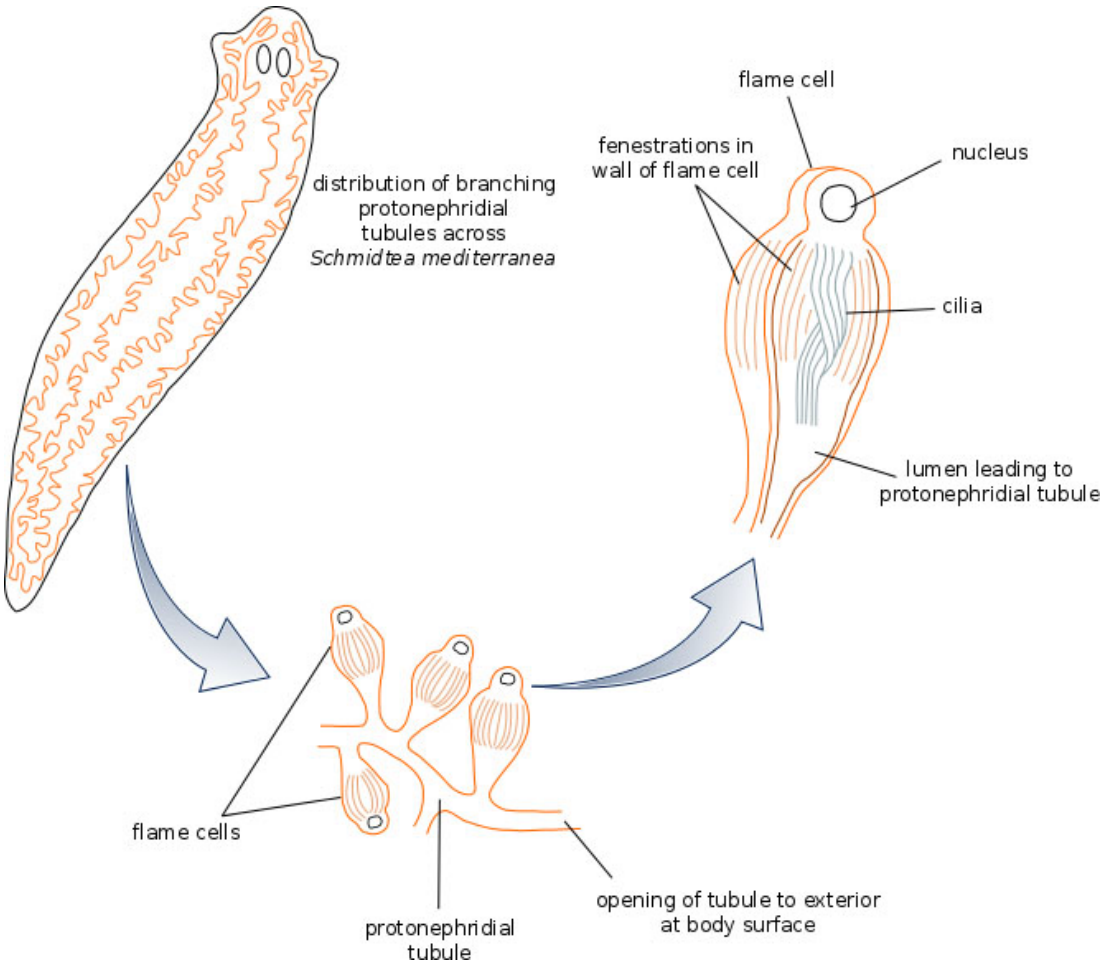


Figure 1.10. Protonephridial tubules distributed across the body of *S. mediterranea*. Diagram modified from Campbell & Reece, (2004)[42]. Tubules are capped with so-called 'flame cells', which function as the site of ultrafiltration. Cilia beat within the tubule lumen to draw body fluid in through fenestrations in the wall of the flame cell. Filtrate then passes into the canal region of the protonephridial tubule for modification prior to excretion.

Protonephridia themselves have been separated into three classes, determined by differences in the morphology of their terminal structure[43]:

1) flame cells: a composite structure composed of the flagellated terminal cap cell and an associated cytoplasmic tubule cell
2) flame bulbs: a single cell structure composed of a blindly ending tube of cytoplasm containing flagella
3) solenocytes: long tubular cells with a nucleus in the cap of the cytoplasm ad a single flagellum lying in the lumen of the tubule.

The distribution of these structures is well correlated with different protostome phyla. Flame cells as a composite structure are found principally in the Platyhelminthes[39,40,43,44]; flame bulbs are most commonly associated with Rotifera[45]; and flagellated solenocytes are found in Priapulida[46], Gastrotricha[43,47], and some adult members of the Annelida[48]. Furthermore, in some coelomate groups which possess metanephridia as adults – Annelida, Mollusca, and Phoronida – protonephridia of varying morphology are found during the larval stage. These either degenerate during metamorphosis or contribute to the adult metanephridial system. Interestingly, the ontogenic relationship between larval protonephridia and adult metanephridia is not consistent between taxa.[49]

1.3.6.2 Metanephridia

Metanephridia are found exclusively in coelomate organisms, where the coelomic cavity acts as a temporary storage compartment for the primary filtrate (Figure 1.9). The term 'metanephridium' is used strictly to describe the ciliated funnel that opens into the coelomic cavity[41]. This serves only as the site of secretion and reabsorption, and leads to the tubule through which waste is eventually excreted. The vertebrate nephron – a 'specialised' metanephridial system – differs from invertebrate metanephridial systems due to its increased structural compartmentalisation. In typical invertebrates, such as annelids, the primary filtrate passes into a generalised coelomic

cavity with multiple functions, including internal transport, hydrostatic integrity, and gamete maturation. Conversely, in the vertebrate nephron, primary filtrate passes into the nephrocoel, or Bowman's capsule, which has the sole function of collecting fluid before it passes to the renal tubule.

The site of ultrafiltration itself is distinct from the metanephridium funnel, and is mediated by specialised epithelial cells called podocytes (described in section 4.1.2.1), found lining the blood vessels in the excretory system. Podocytes have been reported in members of all phyla which possess metanephridial systems, including the Annelida[50], Mollusca[51], and Phoronida[52] within the protostomes; and the Vertebrata, Hemichordata[53], Echinodermata[54], and Cephalochordata[55] within the deuterostomes. In the Arthropoda, podocyte-like cells called nephrocytes mediate ultrafiltration in a modified metanephridial system.

1.3.6.3 Unique ultrafiltratory systems

Despite the widespread distribution of protonephridial and metanephridial systems, not all excretory systems in the Bilateria can be classified into one of these two structural types.

The cephalochordates present an example of a novel, modified metanephridial structure. Excretory systems in amphioxus are two-fold, found as numerous branchial nephridia and as the anterior-most unpaired Hatschek's nephridium, which develop alongside the larval mouth[55-57]. Specialised filtratory cells, 'cryptopodocytes', derive from the subchordalcoelom (in branchial nephridia) and the coelomic epithelium (in Hatschek's nephridium). As in podocytes, the basal epithelial components of the filtration cells form pedicels, which surround the glomerular plexus – a component of the blood vascular system. Initial filtration is believed to occur in the vascular system, across the podocytes and into the coelom. Uniquely, the cryptopodocytes are ciliated at their apical end, surrounded by a ring of 10 microvilli which traverse the coelomic cavity and project into the nephridial canal[55]. The position and solenocyte-like structure of the cryptopodocytes

suggest that the primary filtrate is subsequently conveyed along a canal, induced by the beating of the cryptopodocyte cilia, and into the nephridial duct. The microvilli are also thought to act as a secondary filtration barrier, so that the primary filtrate is secondarily filtered during its passage to the nephridial duct[56]. Despite similarities between the ciliated portion of the cryptopodocytes with protonephridial terminal cells, it is more likely that these cells are apomorphies of amphioxus, and function as a unique ultrafiltratory site[58]. A possible explanation for the presence of Hatschek's nephridium is that the larvae of amphioxus develop precociously, and hence require a functioning filtratory and excretory system to be in place prior to the final differentiation of the branchial region of the trunk.

In the Nematoda and Tunicata, there is no indication for the presence of any ultrafiltratory apparatus. A uniquely simplified excretory-secretory system is present across the Nematoda, which can be broadly divided into two different classes. In most marine nematode species, where water balance need not be so closely regulated, a single large gland – 'renette' – cell, leads directly to an external duct and appears to function in excretion[59]. In many terrestrial species, a branched canal cell, necessary for osmoregulation, is found in addition to a pore cell. In *Caenorhabditis elegans*, three distinct elements (canal, duct and pore) make up a tubular organ with a continuous lumen, but variation in the number and location of the canal cells, and degree of fusion between duct and pore cells, is reported across the Nematoda[60].

The urochordates represent the only deuterostome members to lack obvious tubular excretory organs or any ultrafiltratory specialisation. Instead, ammonia is lost via direct diffusion across their tissues, although some ascidians appear to sequester nitrogenous waste as crystallised uric acid in specialised vesicular cells[61]. In the case of the latter, it appears that this waste remains stored in the cells throughout the lifetime of the animal. Osmoregulation and a degree of solute/fluid balancing appear to occur independently though the neural gland complex, with uptake of molecules across the neural gland reportedly mediated by ciliary beating[62]. This unique

situation found in the tunicates is at odds with the ultrafiltratory and metanephridial systems present across the rest of the deuterostomes (albeit modified in amphioxus). It has been suggested that their putative osmoregulatory system has undergone extensive secondary reduction during their evolution, and hence no longer resembles the nephridial systems found in other bilaterian members.

Some phyla appear to lack excretory systems altogether: no evidence of any ultrafiltratory or excretory specialisation has been found for any members of the Chaetognatha or Bryozoa. One family within the Acanthocephala, the Oligacanthorhynchidae (Archiacanthocephala), have modified protonephridia, which form part of a uro-genital system[63], but excretory systems have not been identified in any other family within this phylum. Similarly, in the Cycliophora, protonephridia have only been reported in the larvae of *Symbion pandora*[64].

1.3.7 Degree of homology of excretory systems

A number of different questions can be addressed regarding the homology of nephridial systems:

1) Are different protonephridial systems homologous to each other?
2) Are different metanephridial systems homologous to each other?
3) Are complete nephridial systems across the Bilateria (comprising a site of ultrafiltration, a tubule element for filtrate modification, and an excretory pore) homologous to one another?
4) Lastly, can we investigate the homology of discrete aspects of nephridial systems: the site of ultrafiltration itself, - that is, protonephridial terminal structures, podocytes, and nephrocytes; and the canal cells and tubules necessary for osmoregulation.

The structure of protonephridia varies widely across the phyla in which they are found, particularly with regards to the terminal structure.

Furthermore, the number of protonephridia present within an organism is not consistent. Such morphological disparity has been interpreted as evidence for the independent evolution of protonephridia in different taxa[55]. However, there are clear similarities in the generalised tripartite structure of protonephridia, and there is now a general acceptance for the homology of protonephridia[71]. At least for members of the Lophotrochozoa, the presence of a single pair of anterior larval protonephridia is assumed to be the plesiomorphic condition, with each monociliated protonephridium deriving from the ectoderm and comprising three cells: one terminal, one duct and one nephropore[61,74]. In this model, modification of the ancestral protonephridial system – for example the presence of more than one terminal or canal cell, or an increased number of cilia in the terminal cell – could have been necessary for increased osmotic regulation in a freshwater environment[75]. Protonephridial systems have, to date, been found exclusively in the Ecdysozoa and Lophotrochozoa: accordingly, it can be hypothesised that protonephridial systems were first present in the last common ancestor of the protostomes.

Metanephridial systems have proved more difficult to assign homology to, with a degree of support for their convergent evolution. Fundamental developmental differences in the ontogenic relationship between larval protonephridia and adult metanephridia in Annelida, Mollusca and Phoronida have been cited as evidence for this theory[74,76]. However, as the development of nephridia is not consistent even within the same phylum – for example Mollusca – this is perhaps not a reliable criterion upon which to assign homology.

Identifying a metanephridial system requires the presence of a ciliated tubule opening into the coelomic cavity of the organism. As the metanephridial tunnel is concerned solely with osmoregulation and modification of the primary filtrate, the presence of specialised ultrafiltratory cells – podocytes – in the epithelia of blood vessels have also been reported in all taxa believed to possess metanephridia. As a result, we can say that for

a species to be confidently described as having metanephridia, both specialised epithelial cells (podocytes) and an open ciliated funnel must be present. Furthermore, both of these elements are necessary in order to fulfill the requirements of a nephridial system: that is, ultrafiltration and osmoregulation.

An intriguing anomaly in this classification is observed in members of the Arthropoda. Although the site of modification – the nephrocytes – appear to be homologous to the podocytes, the insects and some chelicerate species are commonly described as having a 'modified' metanephridial system in the form of blind-ending Malpighian tubules[11]. Although they carry out an equivalent function to the tubule element of other metanephridial systems, the fact that they are blind-ending contradicts the 'strict' definition of an open, ciliated metanephridia, which could not only call into question this 'metanephridial' classification, but also the homology of metanephridial systems where they are found.

1.3.8 Functional vs. phylogenetic distribution

Trying to elucidate the homology of excretory systems – and of metanephridia in particular – is perhaps not aided by the inconsistent distribution of different nephridial systems across the Bilateria (Figure 1.11). The presence of a coelomic cavity is mandatory for the occurrence of metanephridia. They are the only nephridial system found in the Deuterostomia (to the exclusion of the urochordates), but are also found in many protostome groups. Despite this, not all organisms with a coelomic cavity possess metanephridia: this is true for Rotifera, some adult members of the Annelida and, to a certain extent, the Nemertea, which all have protonephridia in the adult form. Most larval Phoronida[65], Annelida[66] and Mollusca[67] have protonephridia, but develop metanephridia as adults after undergoing metamorphosis.

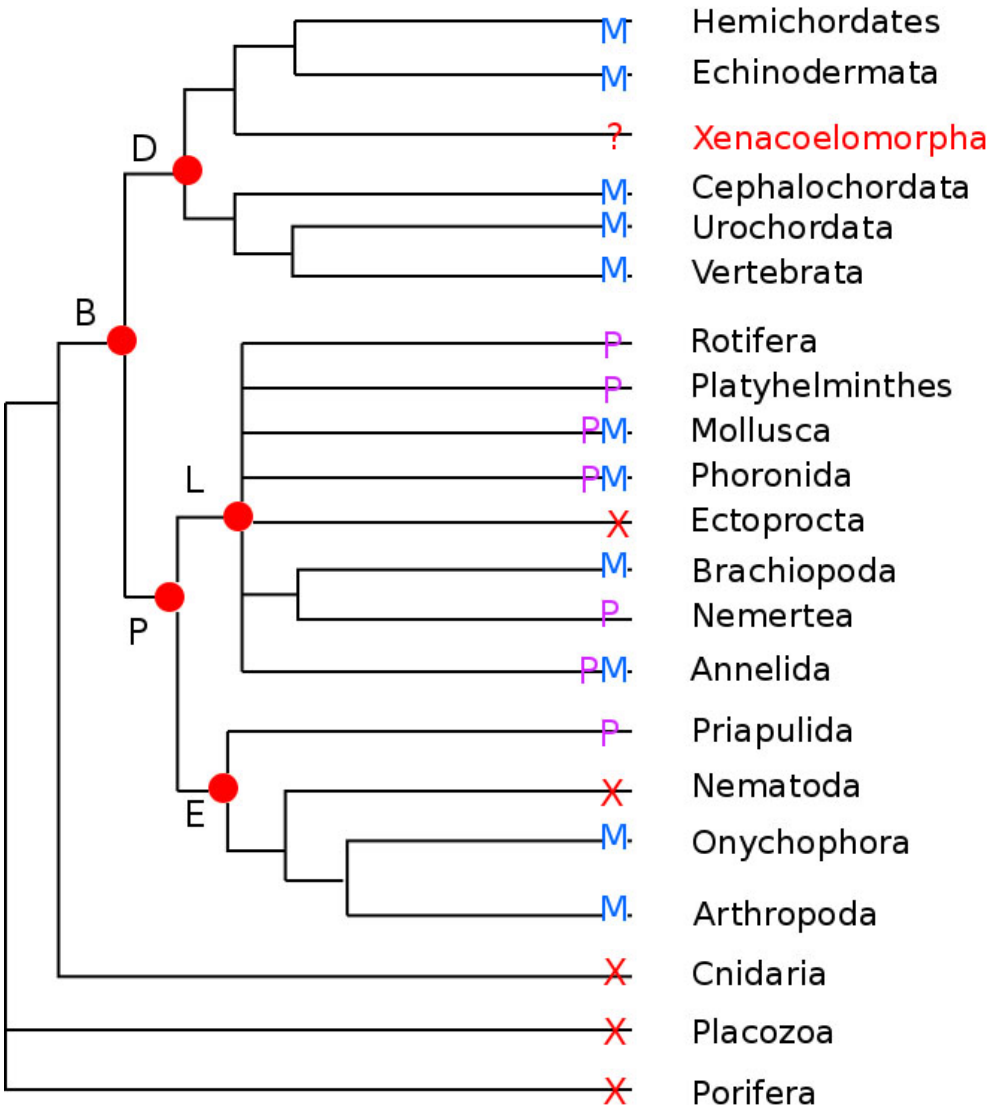


Figure 1.11. Distribution of protonephridial and metanephridial systems across the Bilateria. Purple 'P' designates presence of protonephridia in a lineage; blue 'M' designates presence of metanephridia. Red X indicates assumed absence of a conventional nephridial system. Groupings at nodes: D = Deuterostomia; B = Bilateria; L = Lophotrochozoa; P = Protostomia; E = Ecdysozoa.

Early discussion on nephridia by Goodrich (1934, 1945)[68,69], a proponent of the gonocoel theory, centered around the argument that according to the phylogenetic distribution of excretory systems, protonephridia ('first nephridia') must be primitive to metanephridia. According to this, protonephridia represent the ancestral state, and were able to differentiate into metanephridia depending on expansion of the coelomic cavity.

Based on the body size and vascularisation of different taxa, a functional argument for the presence of different nephridial systems has also been proposed[40]. The underlying differences between protonephridia and metanephridia are the location of the filtration site (that is, of vascular fluid across endothelial cells into the coelom in metanephridial systems, and of extracellular fluid into variable compartments in protonephridia), and the mechanism for driving passive ultrafiltration (muscle-mediated in metanephridia, and cilia-mediated in protonephridia). Consequently, it has been hypothesised that the type of nephridial system present in certain taxa is dependent not on evolutionary history, but on the presence or absence of a coelomic cavity and a circulatory system, which is mandatory for the occurrence of metanephridia. In such a model, protonephridia are not primitive structures, but develop owing to the constraints of a small body size and absence of a blood vascular system. Animals with protonephridia in their larval form that develop metanephridia following metamorphosis are cited as evidence for these constraints– that is, small larval body size and a lack of blood vascular system – on structure. Where protonephridial systems persist into the adult form in some polychaetes, this is attributed to the absence of an adult circulatory system.

The structure of nephridial systems must also be intrinsically linked to their functional requirements. Bilaterians occupy a diverse range of habitats, and different environments require different osmoregulatory capacities. Animals that have a marine habitat encounter less variability in their osmotic environment than animals living terrestrially or in freshwater. As we can see in marine nematode species and in *S. mediterranea*, this seems to result in a

reduced demand for water reabsorption, and the loss of canal cells (nematodes)[60] or intermediate tubule elements (*S. mediterranea*)[37].

1.3.9 Investigating the origin of nephridial systems: the Xenacoelomorpha

The evolution of organs for excretion and osmoregulation in the Bilateria poses an intriguing system in which to investigate the homology of a common function carried out by very diverse structures. Whilst nephridia have largely been compared from a morphological perspective, recent publications have focused on investigating common molecular features in some aspects of nephridial systems: in particular at the site of ultrafiltration, and in the tubule elements required for modification and absorption. The identification of common transcription factors and structural, membrane bound and transporter proteins could prove useful not just for determining the homology of nephridial systems, but also for providing ways to search for and identify possible precursors of nephridia in non-bilaterian animals. Putative genetic markers of nephridia may also help us to identify the sites of filtration and excretion in bilaterians in which such features have not been identified.

The pivotal phylogenetic position of the Xenacoelomorpha therefore makes them an interesting group in which to investigate the origin and evolutionary history of nephridia. Given that they are assumed to lack any excretory specialisation, identifying the presence of such nephridial genes in xenacoelomorphs could also imply a degree of organisational complexity that has previously not been anticipated for this phylum.

Four possible scenarios can be considered with regards the Xenacoelomorpha and the distribution of nephridial structures. Firstly we can consider Xenacoelomorpha in a basal bilaterian position, between the diploblasts and the rest of the triploblastic bilaterians (Figure 1.6). In line with the main protostome and deuterostome grouping being termed the Nephrozoa, this would imply that nephridia evolved after the divergence of

the Xenacoelomorpha from the rest of the Bilateria, and the lack of nephridia in xenacoelomorphs would therefore be a primary absence. Should putative ultrafiltratory structures be found in Xenacoelomorpha members in this hypothetically basal position, the Nephrozoa grouping would no longer be exclusively applicable to the protostomes and deuterostomes.

Conversely, if Xenacoelomorpha are considered to be deuterostomes lacking nephridial structures, then the Nephrozoa grouping would no longer encompass all protostomes and deuterostomes, and it could be hypothesised that the lack of nephridia in Xenacoelomorpha is a secondary absence (Figure 1.6). Lastly, if Xenacoelomorpha members are deuterostomes with ultrafiltratory structures, then the Nephrozoa grouping would remain applicable to all protostomes and deuterostomes.

As there is no consensus for either the homology of excretory systems within the Bilateria, or for the phylogenetic position of Xenacoelomorpha, both of these questions remain intriguing for our understanding of Urbilateria and the evolution of bilaterian novelties. In the last part of this introduction I will outline the objectives of my thesis concerning the evolution and morphology of the Xenacoelomorpha: firstly with a focus on visualising the expression of molecular markers of excretory systems using *in situ* hybridisation and immunohistochemistry; and secondly, using novel RNA-Seq approaches to increase our understanding of cell-type specialisation in *Xenoturbella bocki.*

1.4 Objectives of Thesis

1.4.1 Investigating the presence or absence of ultrafiltratory systems in the Xenacoelomorpha

As described, the disparity in morphology of excretory systems across the Bilateria meant that until relatively recently, they were most commonly defined not by a common structure, but by their common 'tripartite' function of

filtration, modification, and excretion. More recent molecular and evo-devo studies, specifically at the site of ultrafiltration and in tubule elements, have in fact revealed a degree of molecular conservation between distantly related taxa that was perhaps surprising (see section 4.1.3). Consequently, we can identify putative molecular markers of aspects of excretory systems, which are consistently associated with a specific function in homeostasis or excretion.

The first objective of my thesis is to investigate the expression and function of so-called ultrafiltratory genes in members of the Xenacoelomorpha. That is, to use molecular protocols and gene visualisation approaches (*in situ* hybridisation and immunohistochemistry) to establish where these genes are expressed, and whether the xenacoelomorphs have cells that are putatively homologous to the ultrafiltratory cells found in other bilaterians.

1.4.2 Establishing molecular protocols to investigate ultrafiltratory capacity

Within the context of molecular studies, certain taxa have become well recognised as 'model organisms': species which can be cultured or bred in the lab, for which reliable protocols have been established, and for which we have a comprehensive set of molecular data available. Historically, a number of model organisms have been frequently used in evo-devo studies, covering many of the branches of the bilaterian tree. These include the fruit fly *D. melanogaster* (Hexapoda, Arthropoda), the flatworm *S. mediterranea* (Rhabditophora, Platyhelminthes), the nematode *C. elegans* (Rhabditida, Nematoda), and many vertebrate species including the mouse (*Mus musculus*) and zebrafish (*Danio rerio*). Whilst lab studies involving these taxa have contributed hugely to our understanding of development and gene function, they represent only a limited subset of morphological and developmental diversity across the Bilateria. As evo-devo studies expand to include other, traditionally non-model species, it is evident that inferring

findings in one species to be true across long evolutionary distances, or even within a given phylum, is often not valid. Given the uniquely simple morphology of the Xenacoelomorpha, and the limited knowledge we have of their biology (which is particularly true for *Xenoturbella* species), establishing molecular protocols is necessary to better-understand their development and organisation, and specifically to investigate the function of so-called ultrafiltratory genes.

Selected molecular-lab based approaches such as *in situ* hybridisation, immunohistochemistry, and RNAi have been used successfully in some acoelomorph species. Of these, *Isodiametra pulchra* and *Hofstenia miamia* represent perhaps the closest we have to model acoels: they can be maintained long-term in culture, and have accessible embryos, juveniles, and adults for experimentation. Gene expression studies in *I. pulchra* have contributed to our understanding of their nervous system, stem cell system, and germ layer specification, amongst others[70-72]. RNAi in *H. miamia* has been used to understand mechanisms underlying regeneration in the Acoelomorpha[29]. However, the acoels comprise a wide diversity of species. Consequently, expanding the focus of evo-devo studies to include other representatives from the Acoelomorpha is necessary to inform our understanding of the development and molecular pathways of specification in this group. Closely linked to this is the need for increased molecular data: reliable transcriptomic or genomic data is a fundamental starting point for laboratory protocols.

To date, no molecular protocols have been established in *Xenoturbella*. Until very recently, *X. bocki* was the lone representative of this group, and the difficulty associated with collecting these animals, which live buried in the mud at a depth of ~100m in a fjord on the west coast of Sweden, has been a significant barrier to their use in lab-based techniques. We also know very little about their life cycle or ecology, and embryological or developmental studies in this species are challenging. Although the discovery of four new *Xenoturbella* species increases the molecular data for this genus, they live in even more inaccessible environments than *X. bocki*,

at depths of between 700 and 3700 metres in the Gulf of California and Monterey Canyon[23]. Consequently, *X. bocki* presents the most likely candidate for molecular protocols, and establishing experimental procedures in this taxon would be helpful for our understanding of their body plan and degree of structural specification. Furthermore, their very simple morphology and apparent lack of organ systems means that expression studies for genes commonly associated with certain systems or functions might inform our understanding of ancestral gene function - with a focus in this thesis on ultrafiltratory-related genes.

1.4.3 Using novel RNA-Seq technologies (Tomoseq and single cell sequencing) in *Xenoturbella bocki*

RNA-Seq technologies have been refined in recent years to be compatible with low input concentrations of RNA, and we can now generate differential transcriptomic data from single cells or small tissue sections. In conjunction with using molecular protocols to investigate ultrafiltratory gene expression, the second objective of my thesis is to use novel low input RNA-Seq approaches to investigate the presence of ultrafiltratory cells, and to understand organisational complexity and gene expression domains in *Xenoturbella bocki*. I first aim to use whole-organism single cell sequencing in *Xenoturbella* in conjunction with molecular approaches to investigate the presence of ultrafiltratory or excretory related cells, and more widely to investigate the degree of cell-type complexity and specificity. Secondly, I aim to use a low input RNA Tomography approach to generate differential transcriptomic data for *Xenoturbella* with a degree of spatial resolution.

1.4.3.1 Cell type diversity and single cell sequencing

As introduced in section 1.2.3, all metazoan animals are defined by the presence of multiple cell types and differentiated tissues. The number of different cell types in an organism has historically been used as a classificatory tool for inferring evolutionary complexity[73]. Diploblastic

lineages, including the cnidarians and ctenophores, amongst others, have widely been considered to lack numerous or complex cell types; increased organisational complexity in the bilaterians is thought to have resulted in an increased number of cell types with defined functions – including, for example, the evolution of nephridial systems. Morphological and histological studies of *Xenoturbella bocki* provide evidence for its very simple body plan and organisation, but we know comparatively little about the complexity of the types of cells or tissues that it possesses.

Although *in situ* hybridisation and immunohistochemistry can be used to investigate genes commonly associated with a certain cell type or tissue (for example, ultrafiltratory genes), these experiments are limited by the number of genes that can be investigated in parallel. For *Xenoturbella* and other non-model organisms where comparatively little is known about their morphology, this presents a challenge to large-scale gene expression visualisation experiments. Whole organism single cell sequencing therefore presents the opportunity for high-throughput investigation of gene expression and cellular complexity in *Xenoturbella bocki*. With this in mind, I aim to use single cell sequencing alongside *in situ* hybridisation with the aim of identifying cells that express genetic markers of ultrafiltration and excretion. Furthermore, single cell sequencing data for this species would offer a significant contribution to our wider understanding of the types and complexity of cells and tissues that it possesses.

1.4.3.2 Spatially resolved transcriptomics

Alongside the single cell sequencing approach, I aim to combine the manual dissection of *Xenoturbella bocki* along a defined body axis with low-input RNA-Seq technologies to investigate differential transcriptomics with a degree of spatial resolution. This approach, termed RNA Tomography, or Tomoseq[74], represents a useful tool for investigating variable gene expression along different body axes of the adult animal. In particular, I aim to investigate the differential expression patterns of genes with well known

and conserved domains of expression in other taxa. For Tomoseq to be reliably implemented in *Xenoturbella bocki*, I hope to establish an efficient RNA extraction protocol for small cryosectioned tissue slices and combine this with an RNA amplification technique that is prone to minimal amplification biases.

1.5 Overview of Thesis

In the following chapter I outline the materials and methods used in this thesis. In Chapter Three I address the phylogenetic question regarding the position of the Xenacoelomorpha by sequencing and describing the mitochondrial genomes from three species of acoel: *Paratomella rubra*, *Isodiametra pulchra*, and *Archaphanostoma ylvae*. Mitochondrial genomes are valuable tools for inferring evolutionary relatedness, and this work increases the mitochondrial data available for the Acoelomorpha, and includes phylogenetic inference using mitochondrial protein-coding gene sequences.

In Chapter Four I address the question of homology of excretory systems in the Bilateria, and specifically the degree of morphological and molecular conservation at the site of ultrafiltration in diverse bilaterian taxa. In particular, I introduce the functional conservation of the key 'ultrafiltratory' genes in nephridial systems, and investigate the presence or absence of these genes in metazoan taxa commonly regarded to lack organs specialised for ultrafiltration and/or excretion, with an emphasis on members of the Xenacoelomorpha.

As has been described, molecular lab techniques in the Xenacoelomorpha are not well established. Gene expression profiling by *in situ* hybridisation has been successful to some extent in *I. pulchra* and *S. roscoffensis*, but this has not been used for a wide set of genes, and the protocols established are not always straightforward to implement. In Chapter Five I outline the gene expression visualisation protocols I have

used in *S. roscoffensis* with the objective of identifying the expression domains of known molecular markers of ultrafiltration. No molecular protocols in *Xenoturbella* have been reported to date. In Chapter Six, I describe the techniques used to implement *in situ* hybridisation and immunohistochemistry in sectioned adult *Xenoturbella* for the first time, with the aim of identifying the expression of genes associated with ultrafiltration.

To complement the molecular lab approaches, two different RNA-Seq methods are carried out with the objective of understanding more about gene expression and cell types in *Xenoturbella*. Whole organism single cell sequencing is described in Chapter Seven; and the spatially resolved transcriptomics approach 'Tomoseq' is described in Chapter Eight. Both of these are novel techniques for *Xenoturbella* and indeed for any member of the Xenacoelomorpha.

Finally, in Chapter 9 I give a general discussion on the findings and the wider implications of the work carried out in this thesis, and discuss the objectives for further investigation into the Xenacoelomorpha.

2 Material and Methods

2.1 Animal Collection and Culture

2.1.1 Acoelomorpha

Paratomella rubra were collected from the beach in Filey, near Scarborough, North Yorkshire, UK. Samples of the top 20cm of sand were taken from the intertidal zone at low tide and topped up with seawater to be transported back to the lab. 7.2% magnesium chloride ($MgCl_2$) was prepared by dissolving 72g of magnesium chloride hexahydrate ($MgCl_2{}^{*}6H_2O$) in H_2O, to a final volume of 1 litre. Approximately 200ml of $MgCl_2$ was added to ~200g of sand in a flask, swirled, and left to settle for 5 minutes. A filter was made by cutting out the central portion of a 50ml falcon tube lid, stretching 40µM nylon mesh across the opening, and screwing the open lid back onto the tube to keep the mesh tightly in place. The bottom of the tube was cut off so that the liquid portion only of the sand/$MgCl_2$ mix could be poured into the falcon tube and through the filter in the lid. Organisms sedated by the $MgCl_2$, and residual particulate and organic matter >40µM, were retained by the filter and washed from the inside of the falcon tube into glass petri dishes using artificial seawater (ASW, 35 ppt). The contents of the petri dish was examined under a dissecting microscope against a white background, to identify specimens of *P. rubra*. Animals for DNA/RNA extraction were transferred in batches into a 1.5ml Eppendorf tube with as little liquid as possible, flash-frozen in liquid nitrogen, and stored at -80°C until required. Animals required for other molecular approaches were fixed in formaldehyde (as described below).

Samples of *Symsagittifera roscoffensis* were provided by the Sérvice Expedition de Modèles Marins Multicellulaires, pat of the Centre de Ressources Biologiques Marines, at the Station Biologique in Roscoff, France. Animals were kept in ASW in large clear Tupperware boxes covered with cling film, in an 18°C incubator on a 12 hour light/dark cycle. Newly laid

eggs were removed to glass petri dishes on a daily basis, and monitored for hatching. When hatched, animals of an approximately equal age were kept in separate petri dishes and fixed over the course of two weeks to provide a range of fixed juvenile animals at different ages (1-2 days post-hatching to two weeks post-hatching). When no more newly-laid eggs could be found, adult animals were fixed (see below), or flash-frozen in batches and stored at -80°C for DNA/RNA extraction.

Cultured *Isodiametra pulchra* were provided by Bernhard Egger. Animals were kept in glass petri dishes with nutrient-enriched f/2 artificial sea-water[75] and fed ad libitum with the diatom *Nitzschia curvilineata*. Animals were kept at 20°C on a 14 hour/10 hour light/dark cycle.

2.1.2 *Xenoturbella bocki*

Specimens of *Xenoturbella bocki* were collected at Gullmarsfjorden on the west coast of Sweden. A Warén dredge aboard the Oscar von Sydow research vessel (Sven Loven Centre, Kristineberg, Gothenburg University) was used to collect soft mud at a depth of ~100m. Mud was subsequently filtered through two sieves of decreasing diameter, so that larger organisms were removed first, and *Xenoturbella* (of approximately 50-100mm) were retained on the mesh of the second sieve. The contents of the second sieve were washed into large shallow trays, and *Xenoturbella* specimens identified by eye. Animals were kept unfed at 4°C, in sealed falcon tubes, with ASW replaced once a week.

2.2 Animal Fixation

2.2.1 Acoelomorpha

S. roscoffensis, P. rubra and *I. pulchra* adults, and *S. roscoffensis* juveniles were fixed for whole-mount *in situ* hybridisation and immunohistochemistry.

Live animals were anaethetised for 30 min in 7.2% $MgCl_2$ (prepared as described above) on ice. 4% paraformaldehyde (PFA) was freshly prepared with 1x phosphate-buffered saline (PBS). $MgCl_2$ was exchanged with PFA over two five-minute washes, before an hour-long fixation in PFA at 4°C. After several quick washes in 1x PBS, animals were washed for ten minutes each into progressively increased concentrations of methanol (MeOH) (25% MeOH; 50% MeOH; 75% MeOH, all diluted with 1x PBS), and stored in 100% MeOH at -20°C.

2.2.2 *Xenoturbella bocki*

Animals required for RNA extraction were placed directly into 12ml falcon tubes of RNALater (Ambion). Animals required for use in the Tomoseq protocol were washed progressively on ice from ASW into 3.6% $MgCl_2$ (prepared as a 1:1 dilution in deionised water of the standard 7.2% $MgCl_2$ solution described above) and anaethetised in the dark at 4°C for approximately 4 hours. Animals were photographed prior to total loss of movement, and the anterior and posterior ends described. Fully anaethetised animals were carefully placed directly into 12ml falcon tubes of RNALater solution. Tubes were wrapped in aluminium foil and stored at room temperature for one week, before being stored at 4°C until required.

Animals required for *in situ* hybridisation and immunohistochemistry were washed progressively on ice from ASW into 3.6% $MgCl_2$ and anaethetised in the dark at 4°C for approximately 4 hours. Prior to complete loss of movement, the orientation of the animals was described (anterior/posterior; left/right; and dorsal/ventral axis) and individuals were photographed. Fully anaethetised animals were placed onto blotting paper on top of frozen fixative (freshly prepared 4% PFA in 1x PBS) and left overnight (~12 hours) in the dark at 4°C to fix. The following morning, animals were washed several times in 1x PBS and then washed into progressively increased concentrations of methanol (MeOH) (25% MeOH; 50% MeOH;

75% MeOH, all diluted in 1x PBS) and stored in 100% MeOH at -20°C prior to use.

2.3 DNA/RNA Extraction

2.3.1 Acoelomorpha DNA Extraction

Two different protocols were used to extract DNA from the acoels *P. rubra*, *S. roscoffensis* and *I. pulchra*.

2.3.1.1 QIAmp DNA Mini Kit (Qiagen)

For *P. rubra* and *S. roscoffensis*, animal samples were removed from -80°C storage and left to thaw at room temperature. For *I. pulchra*, DNA was extracted from pooled samples of ~50 live animals.

180µl Buffer ATL was added, at room temperature, to the tube containing the specimens. 20µl of Proteinase K (20mg/ml) was added to the sample, and mixed thoroughly by pulse-vortexing. Samples were incubated overnight at 56°C in a heat block.

Following overnight incubation, 200µl Buffer AL was added to the sample, which was pulse-vortexed on a low speed for ten seconds. For all samples, the formation of a white precipitate was observed, which dispersed on re-incubation at 70°C with occasional pulse-vortexing over ten minutes. 200µl of chilled 100% ethanol was added to the sample, mixed by pulse-vortexing for 15 seconds, briefly spun down in a bench-top mini centrifuge and incubated at room temperature for five minutes. The entire lysate was applied to a QIAmp Mini spin column and centrifuged at 6,000 rpm for one minute. Subsequent column wash steps were carried out as per the manufacturer-recommended protocol. Finally, the column was placed in a clean 1.5ml LoBind Eppendorf tube, and 30µl nuclease-free water applied to the column membrane. The column was incubated at room temperature for

one hour to maximise yield, and centrifuged at 13,000 rpm for three minutes. Final sample concentration was measured using a NanoDrop.

2.3.1.2 Phenol/Chloroform extraction (P. rubra)

Where fewer numbers of pooled animals were being used for DNA extraction, a phenol/chloroform approach was used to minimise loss of material on the column.

500µl of fresh SDS lysis buffer and 10µl Proteinase K was added to the sample (<20 animals each time) in a 1.5ml Eppendorf tube. Samples were pipetted up-and-down and pulse-vortexed, and incubated at 50°C for one hour, with occasional pulse-vortexing, to ensure thorough lysis of the tissue.

240µl of phenol/chloroform/isoamyl alcohol was added to the lysed sample, and mixed thoroughly by inverting the tube continuously for 20 seconds. The sample was then centrifuged at 13,000 rpm at room temperature for five minutes. The upper aqueous phase, containing the DNA, was transferred into a fresh 1.5ml tube. To this, 2µl glycogen (equal to 10µg) and a volume of sodium acetate (NH_4OAc, 3M) equal to one tenth of the volume of the aqueous phase was added, and mixed by pulse-vortexing. 2.5x the new solution volume (including the addition of sodium acetate) of ice-cold 100% EtOH was added to the solution, which was mixed by pulse-vortexing and left to incubate overnight at -20°C.

The following day, the sample was centrifuged at 4°C for one hour at 13,000 rpm, and the supernatant removed carefully to leave the pelleted DNA, visible as an opaque white smear. The pellet was washed three times in freshly prepared 70% EtOH, centrifuging at 4°C for three minutes at 13,000 rpm each wash. After the final wash of EtOH was removed, the pellet was left to air dry for 15 minutes, until no visible droplets of EtOH were

visible. The dried pellet was resuspended in 35µl of nuclease-free water, and quantified using a NanoDrop.

2.3.2 Acoelomorpha RNA Extraction (*S. roscoffensis*)

300µl of Trizol (Invitrogen) was added to pooled samples of 40-50 animals in a 1.5ml Eppendorf tube. Tissue was dissoluted by pulse-vortexing and crushing and mashing the tissue using an RNase-free disposable pestle. Once the tissue was completely dissoluted, a further 700µl of Trizol was added, to a total of 1ml, and the solution pulse-vortexed. 200µl of chloroform (0.2x volume of Trizol) was added to the solution, and the tube inverted vigorously for one minute until a homogenous opaque pink colour. After standing at room temperature for three minutes, the sample was centrifuged at 4°C for 20 minutes at 13,000 rpm.

The upper aqueous phase, containing the RNA, was removed into a clean 1.5ml Eppendorf tube. To this, 2µl glycogen (equal to 10ug) and 500µl of 100% isopropanol (0.5x volume of Trizol) was added, and mixed. Samples were incubated at room temperature for ten minutes and stored overnight at -20°C.

The following day, samples were centrifuged for one hour at 13,000 rpm at 4°C. The supernatant was carefully removed and discarded, and the RNA pellet washed three times in 80% EtOH, centrifuged for 20 minutes at 4°C at each wash step. After the final wash, the RNA pellet was air-dried for ten minutes under a lamp, whilst being monitored carefully to prevent over-drying. Once dry, the pellet was resuspended in 30µl nuclease-free water. Final RNA concentration was quantified using a NanoDrop and an aliquot of the sample run on a 1% agarose gel to verify successful extraction.

2.3.3 *Xenoturbella bocki* total RNA Extraction

Animals from which total RNA was to be extracted were placed directly into RNALater for a minimum of one week. RNA extraction was carried out using a custom phenol/chloroform and column extraction hybrid protocol, as outlined in 2.11.2.

2.4 Sequencing and verifying acoel mitochondrial genomes

2.4.1 Polymerase Chain Reaction

PCRs were carried out using the Expand Long-Range PCR Kit (Roche) in 50µl reaction set-up:

Reagent	Volume
Nuclease-free water	38µl
Expand long-range buffer	10µl
dNTPs (equimolar solution of 10mM per dNTP)	2.5µl
Forward primer (20 pmol/µl)	0.75µl
Reverse primer (20 pmol/µl)	0.75µl
DMSO	1.5µl
Template DNA	1µl
Expand long-range enzyme mix	0.7µl

General cycling protocol was: 92°C for two minutes; 15 cycles of 92°C for 10 sec, 57°C for 15 sec, 68°C at initial elongation time (calculated as one min per 1000 base pairs to be amplified); two cycles each of: 92°C for 10 sec, 57°C for 15 sec, 68°C at 40 sec longer than initial elongation time, repeated at increasing 40 sec intervals for a further 14 cycles; a final elongation stage at 68°C for seven min and a 4°C 'hold' stage. Where PCRs were not successful using this protocol, they were repeated using the Q5 High-Fidelity PCR Kit (New England Biolabs), following manufacturer recommendations for a 25µl reaction:

Reagent	Volume
Q5 High Fidelity 2x Master Mix	12.5µl
Nuclease-free water	8.75µl
Forward primer (20 pmol/µl)	1.5µl
Reverse primer (20pmol/µl)	1.5µl
Template DNA	1µl

Cycling protocol was: 92°C for two minutes; 40 cycles of 92°C for ten sec, 58°C for 15 sec, 68°C at initial elongation time (approximated as one min per 1000 base pairs to be amplified); and a final elongation stage at 68°C for seven min. Amplified products were visualised on ethidium-bromide stained TAE 0.8% gels. Bands of expected size were purified using the High Pure PCR Product Purification Kit (Roche Applied Sciences) and sent for sequencing by Source BioScience Life Sciences. Only amplifications that resulted in one clear band on the TAE 0.8% agarose gel were sequenced.

2.4.2 Confirming mitochondrial contigs

Three fragments of sequence from the mitochondrial genome of *P.rubra*, of size ~5.8kb, ~4kb, and ~1.2kb, were generated from gDNA assembly. Fragments were verified using a translated nucleotide query BLAST with invertebrate codon usage (blastx NCBI), and their orientation determined by gene annotation in comparison to the published 9.7kb section of the *P. rubra* mitochondrial genome. Primers were designed in conserved gene regions to:

1) Amplify across the 'N-stretches' present in the 5.8kb and 4kb fragments (eight and nine N-stretches respectively, all of arbitrary length 50 base pairs)
2) Cover the whole 1.2kb fragment, with the aim of resolving the two frameshift mutations within the assembled sequence

3) Close the circular mitochondrial genome, by joining the 5.8kb fragment to the 1.2kb fragment; the 1.2kb fragment to the 4kb fragment; and the 4kb fragment to the 5.8kb fragment (see Appendix 1).

Amplification of the fragments joining the 1.2kb fragment to the 4kb fragment and to close the mitochondrial genome using standard PCR cycling were unsuccessful. These were repeated using a touchdown protocol with Expand Long-Range polymerase. Annealing temperature was set at 65°C with decreasing 2°C intervals every two cycles down to 49°C. Initial elongation time was calculated as before, increasing 30 sec every two cycles of the touchdown, with a final 6 cycles at 49°C. This successfully amplified the region joining the 1.2kb fragment to the 4kb fragments, but did not succeed in closing the circular genome. Design of three new forward and reverse primers, tried in all combinations and using variable PCR parameters were unsuccessful in closing the mitochondrial genome. Additional RNA-Seq and DNA genomic sequencing data corroborated the stretches of sequence at either end of the mitochondrial genome but did not aid in closing the circle.

Three mitochondrial contigs of size ~13kb, ~3.5kb and ~1.3kb were identified from *I. pulchra* Trinity transcriptome assembly from total RNA sequencing. A further contig of ~19kb was also identified, covering the entire ~1.3kb and 13kb regions, and ~2.4kb of the 3.5kb sequence. Fragments were verified using blastx, NCBI, as outlined for *P. rubra*, and approximations for the location of protein-coding genes and tRNAs determined using the MITOS mitochondrial genome annotation server. From aligning the sequences, I found that the last (3') 300bp of the 13kb fragment was duplicated in the opposite orientation within the end (3') region of the 3.5kb fragment. Primers were designed to span the 13kb contig in two ~5kb sections, and to join the 13kb contig to the 3.5kb contig in both directions, to close the mitochondrial genome and verify the validity of the duplicated section (detail in Figure 3.3). RNA-seq data for *I. pulchra* were mapped to the

long transcriptome assembly contigs and PCR sequencing results using NextGenMap[76] and visualised using Tablet.

A. ylvae was co-sequenced at very low coverage as part of a *P. rubra* genome sequencing experiment. From an initial paired end assembly of Illumina HiSeq data with the CLC assembly cell software (v.5.0) (https://www.qiagenbioinformatics.com/products/clc-assembly-cell/), the full mitochondrial circle of *A. ylvae* was extracted in a single contig identified with BLAST. This was annotated using MITOS[77] and manual refinement as described for *P. rubra* and *I. pulchra*.

2.4.3 Mitochondrial genome annotation

For *P. rubra*, all sequenced fragments were aligned against the initial scaffold of the 9.7kb published sequence; the 5.8kb, 4kb, and 1.2kb genome assembly sequences; and an additional long genome assembly fragment of length 14,954 base pairs. All contigs and PCR sequencing results were similarly aligned for *I. pulchra*, but without a reference sequence. Alignments were visualised using Mesquite (v3.04) (http://mesquiteproject.org) with invertebrate mitochondrial translated amino acid state colour coding. Where ambiguity remained between PCR sequencing results and genome or transcription assembly fragments, the genome or transcriptome assembly nucleotide sequence was used to establish a final 'consensus' sequence for each mitochondrial genome. This was particularly important for repetitive AT regions – for example, within *P. rubra nad1* - for which PCR sequencing results were inconclusive. In the case of *I. pulchra*, where the validity of the duplicated sections could not be confidently determined, a final consensus sequence of 18, 725 base pairs was resolved.

The region for each protein-coding mitochondrial gene in the *P. rubra*, *I. pulchra* and *A. ylvae* sequences were compared against published mitochondrial genomes using a translated nucleotide query (blastx, NCBI) with NCBI translation table number 5 'invertebrate mitochondrial'. Published

genes from the mitochondrial genomes of the acoels *S. roscoffensis* and *P. rubra* were downloaded from NCBI GenBank and aligned to the new consensus gene sequences of both *P. rubra*, *I. pulchra* and *A. ylvae* to verify the location of protein-coding and ribosomal RNA-encoding genes. The 5' end of protein-coding genes were inferred to start from the first in-frame start codon (ATN), even if this appeared to overlap with the preceding gene. Similarly, the terminal stop codon of protein-coding genes was inferred to be the first in-frame stop codon (TAA, TAG or TGA). If no stop codon was present, a truncated stop codon (T—or TA-) prior to the beginning of the next gene was assumed to be the termination codon, completed by post-transcriptional polyadenylation. tRNA sequences and corresponding putative secondary structures were identified using the MiTFi program within MITOS.

2.4.4 Sequence alignment, phylogenetic analysis, and evolutionary rates

Phylogenetic analysis was performed using a concatenated amino acid alignment of all thirteen protein-coding genes for *P. rubra*, twelve protein-coding genes for *A. ylvae* and eleven protein-coding genes for *I. pulchra*. Phylogenetic inference was carried out in collaboration with Philipp Schiffer (Telford lab). Nucleotide sequences from all three acoel taxa and an additional set of species comprising 54 metazoans, taken from a range of published metazoan mitochondrial genomes representing deuterostomes, protostomes, cnidarians and two species of poriferans as an outgroup were aligned using TranslatorX[78] independently for all genes with the appropriate mitochondrial genetic code set for each taxa included, using ClustalOmega[79] for amino acid alignment (for taxa and Accession Numbers see Appendix 2). Protein alignments were reduced to the most informative residues using trimAl v.1.4.15[80] with standard settings. Regions showing ambiguity in alignment were excluded, so that only blocks of well-aligned sequence were included for analysis.

Neighbour-nets were re-constructed in SplitsTrees v.4[81] to screen the dataset for potentially non-treelike patterns, which could impede phylogenetic analysis. Subsequently, RAxML v. 8.2.9[82] was used to infer maximum likelihood phylogenies from the original and the reduced alignments under the MTZOA model, optimised for mitochondrial gene-derived phylogenies. Bootstrapping was conducted using the 'autoMRE' option in RAxML and the trees visualised with figtree v.1.4 (http://tree.bio.ed.ac.uk/software/figtree/). Bayesian inference was carried out on the same trimmed alignment using PhyloBayes v.4.1[83], also under the MTZOA model. 10 chains were run in parallel, with the tree search stopped at ~3950 trees per chain, with a maximum difference of 0.17 when discarding 400 trees as burnin and sampling every 10th tree per chain.

Geneious software (v.8)[84] was used to calculate sequence differences between the overlapping 9.7kb sections of *P. rubra* mitochondrial genomes originating from the worms sampled in this analysis (Yorkshire, UK) and the previously published study (Barcelona, Spain). For the eight protein-coding genes found within this section, ParaAT (v.2.0)[85] was used to calculate translation alignments and the KaKs calculator (v.1.2)[86] to access substitution rates. Geneious was also used to calculate a difference matrix for the *cox1* barcoding gene of the two *P. rubra* populations in comparison to other acoel *cox1* sequences retrieved from NCBI GenBank.

2.5 BLAST queries and identifying orthologous sequences

Orthologues of vertebrate genes-of-interest were identified in the Xenacoelomorpha by BLAST queries of Xenacoelomorpha transcriptomes (*X. bocki*, *P. rubra*, *I. pulchra*, *S. roscoffensis*). The protein sequences for the *H. sapiens* genes-of-interest were downloaded from NCBI GenBank for use as the query sequences in the BLAST search. These were compared against the Xenacoelomorpha transcriptomes using the tblastn option, which compares a protein (amino acid) query sequence against a nucleotide

sequence database (our transcriptomes), translated in all possible reading frames.

Often, the BLAST queries returned more than one transcript with similarity to the query sequence. This was likely to be owing to several isoforms of the same gene within the transcriptome. The longest transcript with the highest e-value was likely to be the direct orthologue of the original vertebrate gene. To verify this, the 'best' transcript was used as a reciprocal BLAST on NCBI BLAST using the blastx option, which compares a nucleotide sequence, translated in all reading frames, against a protein sequence database.

To verify sequence orthology, the BLAST tool on the NCBI GenBank database was used to search publicly available data for sequences in a number of comparison organisms, using the *H. sapiens* protein sequences as a query. Where possible, taxa were chosen to represent a cross section of diverse bilaterian organisms (Vertebrata, Cephalochordata, Hemichordata, Echinodermata, Arthropoda, Mollusca, Platyhelminthes, Rotifera, Tunicata, Nematoda); to include representatives from the major diploblastic phyla (Cnidaria, Ctenophora, Porifera, Placozoa); and non-metazoan representative species (Amoebozoa, Choanoflagellatae; Filasterea).

2.5.1 Verifying orthology: tree building with RAxML: ultrafiltratory genes and *Xenoturbella* single cell sequencing probe verification

To assign orthology to the sequences identified from initial BLAST queries, maximum likelihood (ML) phylogenetic analysis was carried out using conserved blocks of amino acid alignments in RAxML. For species with an annotated sequence for the ultrafiltratory gene-of–interest on NCBI, this amino acid sequence was taken for inclusion in the alignment. Where no annotated sequence or positively-identifying reciprocal BLAST sequence could be identified for a species, the best-hit sequence (that is, the sequence with the lowest e-value) returned on NCBI was used. For the

Xenacoelomorpha, the longest sequence with identity verified by reciprocal BLAST was used for alignment. For each tree, related protein sequences were included in the alignment to act as out-groups (*Neph1* and *Nephrin*: cell adhesion molecules; *Podocin*: stomatin family proteins; *CD2AP*: SH3 domain proteins; and *ZO-1*: tight junction proteins). *Neph1* and *Nephrin* are both igSF CAM proteins, and were thus included in the same alignment.

For genes identified from *Xenoturbella* single cell sequencing data as highly expressed markers of cell meta-clusters, related genes from across the Metazoa were used to confirm orthology prior to probe synthesis.

In all instances, amino acid sequences were aligned using Mafft[87]. Protein alignments were reduced to the most informative residues by manual trimming: regions showing ambiguity in alignment were excluded so that only blocks of well aligned sequence were included for analysis. Maximum likelihood phylogenies were constructed using RAxML. Bootstrapping was conducted using the 'autoMRE' option in RAxML; the trees visualised with Seaview v.4 and annotated with Inkscape v0.48.1.

2.6 Molecular Cloning and Sequencing

2.6.1 Reverse transcriptase reaction

Reverse transcription of RNA was carried out using reverse transcriptase PCR (RT-PCR), with two different kits: the GeneRacer Superscript III kit (Invitrogen) for *S. roscoffensis* cDNA synthesis, and the Ambion Retroscript kit (Ambion) for *Xenoturbella* cDNA synthesis.

2.6.1.1 GeneRacer SuperScript III (all reagents provided in the GeneRacer kit)

First strand synthesis was carried out with the addition of the following reagents to 10µl *S. roscoffensis* total RNA (approximately equal to 1.2mg RNA):

Reagent	Volume
Random Primers (N_6) (100ng/µl)	1µl
GeneRacer Oligo(dT) Primers (900ng/µl)	1µl
dNTPs (equimolar solution of 10mM per dNTP)	1µl

The mixture was incubated at 80°C for four minutes in a heat block, then placed on ice for one minute before spinning down on a bench-top centrifuge and standing on ice.

The following components were mixed in a separate tube:

Reagent	Volume
DTT (0.1M)	1µl
RNaseH RNase Inhibitor	1µl
5x First Strand Buffer	4µl
SuperScript III Reverse Transcriptase (RT)	1µl

The 7µl mixture from the second tube was added directly into the 13µl mixture in the first tube, and pipetted gently to combine. First strand synthesis was carried out in a thermocycler, annealing at 25°C for five minutes; reverse transcribing at 50°C for one hour and inactivating at 70°C for 15 minutes. A further 1µl of RNase H was added to the solution, and incubated at 37°C for 20 minutes.

The final 21µl of cDNA was stored at -20°C.

Ambion Retroscript Kit

The Ambion Retroscript kit was used for reverse transcription of *Xenoturbella* total RNA.

To two separate tubes the following components were mixed:

Reagent	Volume
Total RNA (approximately equal to 935ng RNA)	2µl
Nuclease-free water	8µl
Tube One: Oligo(dT) Primers	2µl
Tube Two: Random Decamers	2µl

Tubes were incubated at 80°C in a thermocycler for three minutes, removed to ice, briefly spun down in a bench-top centrifuge and placed back on ice.

The following components were added to each tube on ice and mixed gently:

Reagent	Volume
10x RT Buffer	2µl
dNTP mix (equimolar solution of 2.5mM per dNTP)	4µl
RNase Inhibitor	1µl
MMLV-Reverse Transcriptase	1µl

Tubes were incubated at 42°C in a thermocycler for two hours, then heated to 90°C for ten minutes to inactivate the Reverse Transcriptase.

cDNA was stored as separate tubes (Oligo(dT) and random primers) at -20°C, and aliquots of each mixed in equal volumes directly prior to use.

2.6.2 Primer design and sequences

All primer sequences were designed computationally using Primer3 software. Primers were ordered from Eurofins MWG Operon.

2.6.3 Polymerase Chain Reaction (PCR) for amplification of probe sequences

PCR reactions were carried out using the GeneAmp PCR System 2700 (Applied Bioscience). The majority of probe synthesis PCRs were carried out using RedTaq DNA Polymerase (Sigma-Aldrich). Reactions were carried out to a total volume of 25µl, with reagents added in the following volumes:

Reagent	Volume
Nuclease-free water	19.5µl
10x RedTaq PCR Reaction Buffer	2.5µl
dNTPs (equimolar solution of 10mM per dNTP)	0.5µl
cDNA	0.5µl
Forward Primer	1µl
Reverse Primer	1µl
RedTaq DNA Polymerase	0.5µl

General cycling protocol was set up as follows: initial denaturation at 94°C for five minutes; cycles of 30 seconds denaturation at 94°C, annealing at various temperatures for 30 seconds, and elongation at 72°C. This cycle was repeated 40 times, before a final elongation step of seven minutes at 72°C and a 4°C hold stage. Elongation time was approximated as one minute per 1000 bases to be amplified). Annealing temperature was first set as two degrees centigrade lower than the melting temperature (T_m), as recommended by the primer synthesis company (Eurofins MWG Operon). Where this was unsuccessful, PCRs were repeated as a 'touchdown'

protocol, set as 10 cycles annealing at 60°C; 10 cycles annealing at 55°C and 20 cycles annealing at 50°C.

If there was no successful amplification using the touchdown protocol, the PCR reaction was set up again using 0.5µl of the initial failed PCR as template in lieu of cDNA, and new, 'nested' PCR primers, designed within the region spanned by the initial forward and reverse primers. Nested PCRs were run using the touchdown protocol described above.

1µl of the PCR product was visualised on an ethidium-bromide stained TAE 0.8% gel. Where multiple bands were visible on the gel, 22µl of the PCR product was re-run on an ethidium-bromide stained TAE the band-of–interest was excised using a sterile blade on a UV transilluminator plate, and the DNA extracted using the QIAquick Gel Extraction Kit (Qiagen) with manufacture-recommended protocol. Where single bands of the expected size were visible on the gel, PCR products were purified using the QIAquick PCR purification kit (Qiagen). In both instances, DNA was eluted in 10µl nuclease free water.

2.6.4 Cloning PCR Products

Purified PCR products were ligated into pGEM-T Easy (Promega) plasmid vectors by overnight incubation at 4°C of the following mix:

Reagent	Volume
2X Rapid Ligation Buffer	2.5µl
pGEM-T Easy Vector	0.5µl
T4 DNA Ligase	0.5µl
Purified PCR product	1.5µl

TOP10 *E. coli* cells (Invitrogen) were transformed with ligated plasmids. The total ligated product was pipetted into a thawed vial of TOP10 cells, mixed gently and incubated on ice for 30 minutes. Cells were heat-

shocked at 37°C in a heat block for five minutes, and placed back on ice for five minutes. Successful transformation was verified using blue/white screening on LB+ ampicillin plates and colony PCR.

Colony PCR was performed by inoculating the following PCR reagents with a picked white colony. Between three and five distinct white colonies per plate were verified.

Reagent	Volume
Nuclease-free water	20.75µl
10x RedTaq PCR Reaction Buffer	2.5µl
M13 Forward	0.5µl
M13 Reverse	0.5µl
dNTPs (equimolar solution of 10mM per dNTP)	0.5µl
RedTaq DNA Polymerase	0.25µl

Cycling protocol was set up as: initial denaturation for three minutes at 94°C; 25 cycles of: 30 seconds denaturation at 94°C, annealing at 55°C for 30 seconds, elongation at 72°C for an appropriate time (calculated as described previously); final elongation for seven minutes at 72°C. PCR products were run on an ethidium-bromide stained TAE 0.8% gel to check for an insertion of the correct size. Successfully transformed colonies were cultured overnight in 1.5ml Lysogeny Broth (LB), inoculated with ampicillin, by shaking at 225rpm at 37°C.

Bacterial cells were pelleted by transferring the entire LB culture to a 1.5ml Eppendorf tube and centrifuging for 30 minutes at 13,000 rpm. Plasmid DNA was extracted using the QIAprep spin Miniprep kit (Qiagen), following manufacture-recommended protocol. Verification of cloned fragments was carried out by Sanger sequencing of plasmid samples (Source Bioscience LifeSciences), using M13 forward and reverse primers.

Plasmids were stored at -20°C.

2.6.5 DIG-labelled probe synthesis

One clone with the required RNA insert was selected to make probes for each gene-of-interest. Prior to transcription, templates from the plasmids were made using PCR amplification of the target sequence with M13 forward and reverse primers. 50µl PCR reactions were set up, prepared as follows:

Reagent	Volume
Nuclease-free water	36µl
10x RedTaq PCR Reaction Buffer	5µl
Template (Miniprep)	5µl
dNTPs (equimolar solution of 10mM per dNTP)	1µl
M13 Forward	1µl
M13 Reverse	1µl
RedTaq DNA Polymerase	1µl

PCR cycling was set up as: three minutes at 94°C initial denaturation; 40 cycles of: 40 seconds denaturation at 94°C, annealing at 60°C for 40 seconds, and 90 seconds extension at 72°C; followed by a final ten minute extension at 72°C. PCR products were column-purified using the QIAquick PCR purification kit and eluted in 50µl nuclease-free water.

Orientation of fragment insertion into the vector was determined by the Sanger sequencing results, and the appropriate polymerase chosen (either SP6 RNA polymerase and 10x transcription buffer or T7 RNA polymerase and 10x transcription buffer, Roche) to synthesise the antisense strand.

Probe synthesis was carried out by transcription of the purified template in the following volumes:

Reagent	Volume
Purified PCR product	10µl
Nuclease free water	3.5µl
T7 polymerase OR SP6 polymerase	2µl
DIG RNA labelling mix	2µl
T7/SP6 10x transcription buffer	2µl
RNase inhibitor	0.5µl

The above mix was incubated for two hours at 37°C in a heat block. 2µl DNase was added to the mix to degrade the DNA template, and the mix incubated for a further 15 minutes at 37°C. Transcription was terminated by heating to 70°C in a heat block for ten minutes,

To precipitate the labeled RNA probe, sodium acetate (NaOAc, 3M), equivalent to one-tenth of the volume of the transcription reaction (=2.2µl), and 100% EtOH, equivalent to 2.5x the new volume of the mix (=60.5µl), was added to the solution and mixed well. The mixture was incubated overnight at -20°C.

To wash the probe, the solution was centrifuged for at 13,000 rpm for half an hour at 4°C. The supernatant was tipped off, the opaque RNA pellet washed in 1ml of freshly made 70% EtOH, and centrifuged at 4°C for one minute. EtOH was tipped away, and any remaining drops of EtOH pipetted carefully off. The pellet was air-dried on ice for approximately five minutes until all residual EtOH had evaporated. The pellet was resuspended in 50µl nuclease-free water and quantified on a NanoDrop. 2µl of the probe was heated at 95°C for two minutes to denature the RNA, and run on an ethidium-bromide stained 1% agarose gel. Probes were stored at -80°C.

2.7 Animal embedding and sectioning (*Xenoturbella* and *S. roscoffensis*)

Animals were first fixed as described in 2.2.

Fixed animals stored in 100% MeOH were removed from -20°C storage, and MeOH was exchanged with 100% EtOH over a number of quick washes. As much EtOH as possible was removed from the samples, and the animals washed into Histosol. Samples were incubated in Histosol for three 20-minute washes at room temperature. After a total of an hour-long incubation in Histosol, samples were washed into a 50:50 mix of Histosol:paraffin wax. The mix was prepared during the previous washing stages and kept liquid in a hybridisation oven set at 60°C. Two 30-minute washes into the Histosol:wax mix were carried out at room temperature. Finally, samples were transferred into 100% paraffin wax at 60°C, and left overnight in the hybridisation oven. The following day, samples were washed five times in 100% wax at 60°C. Paper molds were filled with 100% wax and left at 60°C for half an hour, before the animals were placed into the mold. Molds were left to solidify at room temperature.

Sectioning was carried out using a microtome, set to the desired section-thickness. Cut sections were 'floated' onto the surface of a 37°C water bath and onto the surface of clean glass slides. Slides with paraffin sections were placed on a warming block at 65°C for 20 minutes to allow the wax to melt slightly and the tissue to bond to the glass. Finished slides were stored at room temperature.

2.8 *In situ* hybridisation protocols

2.8.1 On sectioned *Xenoturbella* and *S. roscoffensis*

2.8.1.1 Slide preparation

All preparatory and prehybridisation stages were carried out in coplin jars treated with RNaseZap (Sigma Aldrich) and rinsed in DEPC-water.

Slides were dewaxed by two five-minute washes in Histoclear, followed by two five-minute washes in 100% EtOH and two-minute washes down an ethanol gradient (90%, 70% and 50% EtOH:DEPC-treated 1x PBS).

Slides were rinsed once in DEPC-treated water, once in PBTw (0.1% Tween-20 in 1x PBS), and once in 2x SSC, diluted from a 20x stock.

2.8.1.2 Hybridisation

250µl hybe buffer with DIG-labelled RNA probe was prepared for each slide. For initial *in situ* set-up, 1µl probe was added to 249µl hybe buffer, but probe concentration varied in subsequent experiments following degree of expression. Hybe buffer containing probe was applied to each slide and incubated overnight under a glass coverslip in a 50% formamide/2x SSC-humidified chamber. Initial hybridisation temperature was set as 60°C, but as with probe concentration, this varied in repeated experiments.

2.8.1.3 Washes

Washing solution composed of 50% formamide, 1x SSC and 0.1% Tween-20 was prepared. The solution was pre-warmed in a water bath for 30 minutes at 42°C, followed by 15 minutes pre-warming to hybridisation temperature. Slides were removed from the humidified chamber using forceps and placed into coplin jars containing the pre-warmed washing solution. Coverslips were allowed to fall off by first dipping the slides briefly in and out of the solution. Coplin jars were left in the water bath set to hybridisation temperature for 30 minutes, and then washed into fresh pre-warmed washing solution for another 30 minutes.

Slides were washed three times in 1x MABT buffer (10x maleic acid buffer diluted 1:10 in Milli-Q H_2O, with 0.1% Tween-20) for ten minutes each time.

2.8.1.4 Visualisation of Probe

Slides were blocked for two hours at room temperature in blocking buffer composed of 1% Roche blocking reagent and 20% sheep serum, diluted in 1x MABT. Anti-digoxigenin-AP Fab fragments (Roche) were diluted 1:1000 in the same blocking buffer. 130µl of the anti-DIG solution was applied to each slide and incubated overnight under a parafilm coverslip in a humidified chamber at room temperature.

The following day, slides were rinsed multiple times in 1x MABT in coplin jars on an orbital shaker. Wash duration progressed in length throughout the course of the day, starting as quick, ten-minute washes for the first three buffer changes and extending to hour-long by the sixth wash. Slides were left overnight in 1x MABT at 4°C.

After overnight incubation, slides were washed twice for ten minutes in NTMT.

The probe was visualised using BM Purple ready-to-use AP substrate (Sigma-Aldrich), in the dark at room temperature. Slides were monitored for colour change throughout the day, and where a longer development step was necessary, kept in the fridge overnight at 4°C. Substrate solution was changed each morning to minimise background staining. Once slides had developed appropriately, they were washed twice in freshly made NMTM solution to stop the colour reaction, and twice in 1x PBS. If no further procedures were required, slides were coverslipped with fluoromount, containing DAPI and imaged. Where subsequent immunostaining was to be carried out, slides were stored in the dark at 4°C in 1x PBS.

2.8.2 *S. roscoffensis* whole-mount PBS-based protocol

2.8.2.1 Animal preparation

Fixed animals were removed from storage at -20°C and transferred into 1.5ml Eppendorf tubes. Approximately 20 adults or 50 juveniles were used each time and all washes were carried out with 1ml volumes for five minutes unless otherwise specified. Animals were rehydrated through ten-minute washes on a bench-top Sunflower mini-shaker down a methanol gradient: once in 60% MeOH:40% 1x PBS with 0.1% Tween-20 detergent (PTw); once in 30% MeOH:70% PTw; and four washes in PTw. Animals were digested for 15 minutes in Proteinase K (0.01mg/ml), diluted in PTw; juveniles were digested in for eight minutes and adults for 15 minutes. No shaker was used during digestion. To stop digestion, 1ml of 2x glycine (2mg/ml) was added to the tube, followed by two further washes in 1x glycine. Animals were washed once in 1% triethanolamine in PTw, followed by two further 1% triethanolamine washes: the first with 3µl acetic anhydride per 1ml wash; and the second with 6µl acetic anhydride per 1ml wash. The triethanolamine and acetic anhydride acetylation step was included to reduce non-specific background staining during the development stage. Finally, animals were washed twice in PTw and re-fixed in 3.7% formaldehyde in PTw for one hour at room temperature, before being washed five times in PTw. During the fourth wash, samples were heated to 80°C for ten minutes to destroy endogenous alkaline phosphatases, which could interfere with probe visualisation later on.

2.8.2.2 Pre-hybridisation

Prior to use, hybe buffer was removed from -20°C storage, thawed to room temperature, and agitated gently to remove any precipitate. As much PTw as possible was removed from the animals, and 500µl hybe buffer added to the tube. Samples were incubated on a shaker for ten minutes at room temperature. This was replaced by 500µl of fresh hybe buffer, and

animals were incubated in a hybridisation oven at hybridisation temperature (variable) overnight.

2.8.2.3 Hybridisation

Probes were diluted in hybe buffer to a total of 500µl. The concentration of probes varied: initial *in situ* hybridisations were carried out with a probe dilution equivalent to 1ng/µl, which was varied in subsequent protocols according to experimental results. Probes were linearised by heating to 80°C for ten minutes. As much hybe buffer as possible was removed from the animals, and the denatured probe added whilst still warm after heating to 80°C. Samples were incubated in a hybridisation oven at hybridisation temperature (variable), over a varying time period (ranging from overnight to seven days).

2.8.2.4 Washes

After an appropriate hybridisation period, the probe was removed and two hybe buffer washes carried out at hybridisation temperature: the first for ten minutes and the second for 40 minutes. 30 minute washes at increasing concentrations of 2x SSC in hybe buffer (75% HB:25% SSC; 50% HB:50% SSC; 25% HB: 75% SSC), followed by 3x 20 minute washes in 0.2% SSC were all carried out at hybridisation temperature. Ten minute washes from 0.05x SSC into PTw (75% SSC/25% PTw; 50% SSC/50% PTw; 25% SSC/ 75% PTw; 100% PTw) were carried out at RT.

2.8.2.5 Visualisation of Probe

Following 5x PTw washes, samples were blocked in 1% blocking buffer (Boehringer-Mannheim), diluted in maleic acid buffer for one hour at room temperature.

The DIG-labelled probe was detected with an alkaline-phosphatase conjugated anti-DIG antibody. Samples were incubated with anti-digoxigenin-AP Fab fragments, diluted in blocking buffer to 1:5000, overnight at 4°C. Following overnight incubation, animals were washed 5x 15 minutes at room temperature in 0.2% Triton-X diluted in 1x PBS; 5x 30 minutes at room temperature in PTw, 3x 10 minutes in fresh alkaline phosphatase (AP) buffer (without $MgCL_2$); and 2x ten minutes in AP buffer with the addition of $MgCL_2$. AP buffer was prepared immediately prior to use.

The probe was visualised with 1% 4-nitroblue-tetrazolium chloride/5-bromo-4-chloro-3-indolyly-phosphate (NBT/BCIP) (Roche), diluted in AP buffer to various different concentrations (1µl/ml to 8µl/ml). The colour reaction was carried out at room temperature in the dark, with specimens moved to 4°C overnight if multiple days staining was required. Specimens were monitored for colour change and stopped once a signal was evident – ranging in time from a few hours to three days. AP substrate solution was changed daily to reduce background staining. To stop colour development, specimens were washed twice in AP buffer (without $MgCl_2$), followed by five quick washes in PTw. Samples were mounted on microscope slides in 70% glycerol.

2.8.3 *S. roscoffensis* whole-mount MABT-based protocol

2.8.3.1 Animal preparation

Animals were removed from -20°C and rehydrated by ten minute washes down a methanol gradient of 70%, 50% and 30% MeOH:DEPC-H_2O in 1.5ml Eppendorf tubes. Animals were washed three times for 15 minutes in maleic acid buffer containing Tween-20 (MABT), prepared as 50ml as follows:

Animals were pre-hybridised in 500µl fresh hybridisation buffer for one hour at hybridisation temperature (set between 50°C and 60°C).

2.8.3.2 Hybridisation

Probes were prepared in fresh hybridisation buffer, diluted to an appropriate concentration (between 0.05ng/µl and 1ng/µl) in a final volume of 500µl. The solution containing the probe was heated to 90°C for 10 minutes; placed on ice for 10 minutes; and heated to 50°C for 10 minutes before being added to the animal sample in Eppendorf tubes. Probes were hybridised at hybridisation temperature for between two days and one week.

2.8.3.3 Washes

Fresh hybridisation buffer was prepared as above and heated to hybridisation temperature. The probe was removed from the samples, and the animals washed into fresh hybridisation buffer for one hour at hybridisation temperature. Animals were washed twice in MABT buffer, once at hybridisation temperature and once at room temperature; three times in 0.1X MABT buffer; and once in 1x MABT buffer.

2.8.3.4 Visualisation of Probe

Following MABT washes, animals were blocked in 500µl 10% goat serum, diluted in MABT, for one hour at room temperature. As described in 2.8.2**,** the DIG-labelled probe was detected with an alkaline-phosphatase conjugated anti-DIG antibody. Samples were incubated with anti-digoxigenin-AP Fab fragments, diluted in 10% goat serum to 1:1000, overnight at 4°C. After overnight incubation, animals were washed five times in MABT buffer at room temperature.

Animals were washed twice in AP buffer, and the probe visualised using the same method as described in the whole-mount PBS protocol. The colour reaction was stopped by washing once in MABT buffer containing 0.05M EDTA, followed by three washes in MABT buffer. Samples were mounted in 70% glycerol.

2.9 Immunohistochemistry (IHC) protocols

2.9.1 Custom polyclonal antibody synthesis

Custom polyclonal antibodies were raised by GenScript, using their Rabbit Polyexpress Premium Protocol for Neph1 and Podocin, and their Mouse PolyTD Polyclonal Antibody Service for Nephrin. Partial fragments of the inferred amino acid sequence for each gene were used as recombinant protein antigens for antibody production. These corresponded to residues 61-400 of the deduced translation of the *SrNephrin* gene; 25-320 of SrNeph1; and 58-288 of SrPodocin-like. Using the GenScript BacPower protocol, the recombinant protein sequences were cloned into an expression vector, transformed into plasmids, and purified. Resulting protein fragments were injected into the appropriate host animal for use as antibody production. Following immunization and test bleeds, the final antiserum was pooled and purified for use in IHC.

2.9.2 IHC on sectioned *Xenoturbella* and *S. roscoffensis*

2.9.2.1 Slide preparation

Slides were washed twice for five minutes in Histoclear; twice for five minutes in 100% EtOH; and then down an ethanol rehydration gradient, for two minutes each time (90%, 70% and 50% EtOH diluted in 1x PBS). Slides were rinsed twice in Milli-Q H_2O. At this stage, slides that had previously been used for *in situ* hybridisation and stored in 1x PBS were included in the protocol.

2.9.2.2 Primary antibody incubation

Slides were washed once for two minutes in 1x PBS with 0.1% Tween-20 detergent (Sigma-Aldrich) (PTw); twice for ten minutes in 1x PBS

with 0.1% Triton-X (Sigma-Aldrich) (PBT); and blocked in 10% sheep serum diluted in PBT for 30 minutes at room temperature.

Primary antibodies were prepared in the blocking solution at the appropriate dilution. The primary antibody working concentration was determined empirically. Commercial anti-Neph1(Kirrel) (Atlas Antibodies, HPA030458) , anti-Podocin(NPHS2) (Atlas Antibodies, HPA049486) and anti-Nephrin(NPHS1) (Novus Biologicals, AF4269) were diluted at 1:50 and 1:100; custom polyclonal antibodies* for anti-SrNeph1, anti-SrNephrin and anti-SrPodocin-like were diluted at 1:100. 130µl of the diluted antibody was applied to each slide and incubated overnight under a parafilm coverslip in a humidified chamber at 4°C.

*for custom antibody preparation see 2.9.1.

2.9.2.3 Secondary antibody incubation

The following day, slides were washed three times in PBT at room temperature. Fluorescent-conjugated secondary antibodies against the appropriate host species of the primary were diluted 1:1000 in blocking solution. For slides incubated with commercial and *S. roscoffensis* custom anti-Neph1 and anti-Podocin primary antibodies, goat anti-rabbit Alexa 568 (Invitrogen) was used. For commercial anti-Nephrin, goat anti-sheep Alexa 568 (Invitrogen) was used, and for *S. roscoffensis* custom anti-Nephrin, goat anti-mouse Alexa 568 (Invitrogen) was used. Animals were incubated with the secondary antibody overnight at 4°C. 130µl of the diluted secondary antibody was applied to the slide and incubated overnight under a parafilm coverslip in a humidified chamber, in the dark, at 4°C.

2.9.2.4 Washing and mounting

Following overnight incubation, slides were washed several times in PBT, once into 1x PBS, and mounted under glass coverslips using fluoromount (Sigma-Aldrich).

2.9.3 Single IHC protocol on whole-mount *S. roscoffensis*

2.9.3.1 Choice of buffer

PBT buffer (1x PBS solution with an appropriate concentration of Triton X-100) was used as a permeabilisation agent in immunohistochemistry experiments. For adult specimens, 5% PBT was used, with a higher concentration of Triton X-100 necessary for permeabilising the membrane of the epidermal cells in fixed adults. A lower concentration of 0.5% Triton X-100 was used for juvenile immunohistochemistry experiments.

2.9.3.2 Animal preparation

Fixed animals were removed from storage and rehydrated from methanol as described in 2.8.2, with the appropriate concentration of PBT used instead of PTw. Following rehydration, animals were washed five times in PBT, and then blocked for two hours in 10% goat serum, diluted in PBT.

2.9.3.3 Primary antibody dilution and incubation

Primary antibody solutions were prepared in the same 10% goat serum blocking solution at the dilutions described in 2.9. As much blocking solution as possible was removed from the animals, and the animals incubated overnight with the diluted antibody at 4°C. The following day, animals were washed up to eight times in PBT: the first four washes every 15 minutes, and all subsequent washes every hour.

2.9.3.4 Secondary antibody dilution and incubation

All following steps were carried out in the dark to prevent photobleaching of the secondary antibody fluorophore. The appropriate

secondary antibody for the host species of the primary antibody was diluted in 10% goat serum. For animals incubated with commercial and *S. roscoffensis* custom anti-Neph1 and anti-Podocin primary antibodies, goat anti-rabbit Alexa 568 was applied in a 1:1000 dilution. For commercial anti-Nephrin, goat anti-sheep Alexa 568 was used, and for *S. roscoffensis* custom anti-Nephrin, goat anti-mouse Alexa 568 was used, both diluted at a concentration of 1:1000. Animals were incubated with the secondary antibody overnight at 4°C.

Negative controls were performed by omitting the primary or the secondary antibody, or in the case of the *S. roscoffensis* custom polyclonal antibodies, by incubating with the host pre-immune serum in place of the primary antibody. Autofluoresence of endosymbionts in adult *S. roscoffensis* was found to be strongest in the green 488 wavelength, and so for all single antibody staining protocols, secondary antibodies conjugated to an Alexa Fluor 568 dye were used, to try and avoid as much autofluorescent signal as possible. For all pre-immune serum experiments, no signal was present.

2.9.3.5 DAPI counterstaining and mounting

Following overnight incubation, animals were washed directly into DAPI solution, diluted in PBT to an equivalent concentration of 300nM from a 5mg/ml stock solution. Animals were washed 10 times in PBT and mounted in 70% glycerol.

2.9.4 Double IHC protocol on whole-mount *S. roscoffensis*

2.9.4.1 Animal preparation and first primary antibody incubation

Animals were rehydrated from MeOH, washed, blocked and incubated with the primary antibody overnight as described in 2.9.3.

2.9.4.2 First secondary antibody incubation

After overnight incubation at 4°C, animals were washed five times in PBT and incubated with the appropriate fluorophore-conjugated secondary antibody, diluted in 10% goat serum in PBT (as described in 2.9.3). Specimens were incubated with the secondary antibody overnight at 4°C.

2.9.4.3 Second primary antibody incubation

Samples were washed five times in PBT and blocked for two hours in 10% goat serum, diluted in PBT. The second primary antibody was diluted to an appropriate concentration as described in 2.9.3. Experiments were designed so that the host animal for the second primary antibody was different from the host of the first primary, so as to prevent conjugation of both fluorescently labelled secondary antibodies to the same primary antibody to give a misinformative overlapping signal. Animals were incubated with the second primary antibody overnight at 4°C.

2.9.4.4 Second secondary antibody incubation

Animals were washed five times in PBT and incubated with the appropriate fluorophore-conjugated secondary antibody, diluted in 10% goat serum in PBT (as described in 2.9.3). Secondary antibodies conjugated to a different fluorophore to those used for the first secondary antibody incubation were used: either Alexa 568 (first secondary) followed by Alexa 488 (second secondary) or Alexa 568 (first secondary) followed by Alexa 488 (second secondary), depending on the experiment. Control experiments were carried out by omitting the first primary or the second primary antibody. Both versions of these experiments failed to yield signal in the channel for which the primary was omitted.

2.9.4.5 DAPI counterstaining and mounting

Animals were washed, counterstained with DAPI and mounted as described in 2.9.3.

2.10 Single Cell Sequencing Protocol

Xenoturbella cell sorting was carried out in conjunction with Sandrine Schmutz and Sophie Nouvalt of the Pasteur Flow Cytometry Platform. Single cell libraries were prepared in collaboration with Baptiste Saudemont and single cell library clustering was carried out in collaboration with Yann Loe Mie, both at the Pasteur Institute.

2.10.1 Dissociation and cell sorting

Xenoturbella adults were collected as described in 2.1.2. Two animals were processed separately to generate 8x 384 well plates for each animal. Dissociation and sorting were done in a single experiment with the same reagents and capture plates from the same batch.

Animals were dissociated by placing them in full strength calcium/magnesium-free and EDTA-free artificial seawater, then transferring them to the same solution with the addition of 0.5ug/ml LiberaseTM (Sigma Aldrich) in one well of a 48-well plate. Dissociation was carried out at room temperature using gelatine-coated pipette tips of decreasing diameter to disrupt the cell suspension over a period of 15 minutes. Dissociation was stopped by the addition of one-tenth volume of 500mM EDTA. The entire volume of dissociated cells was transferred to a 1.5ml Lo-Bind Eppendorf tube and sufficient calcium/magnesium free seawater added to make a total volume of 1.5ml. 3µl of calcein AM from a freshly made 1 µg/µl calcein AM solution was added to the tube to stain the live cells; 2.25µl of propidium iodide at 1µg/µl was added to stain the dead cells. The entire volume was gently pipetted up and down with gelatin-coated tips to ensure

homogenisation of the sample with the calcein and propidium iodide. Cells were placed on ice before sorting.

Cells were sorted into 384-well capture plates containing 2µl of cell lysis solution using a BD FACSAria III fluorescently-activated cell sorting machine. Lysis solution contained 0.2% Triton X-100, RNase inhibitor and barcoded primers. Non-cellular particles were excluded from sorting based on threshold size. Live cells were selected as determined by Calcein positive/PI negative fluorescence, and doublet or multiplets excluded based on FSC-W vs. FSC-H. Immediately after sorting, plates were briefly spun down to ensure complete immersion of cells in the lysis solution, and stored at -80°C. Four empty wells were left in each plate as a control.

2.10.2 Massively Parallel Single-Cell RNA-seq (MARS-seq)

Single cell libraries were prepared as outlined in Jaitin *et al.* (2014)[88]. All 16 384-well plates (a total of 6080 single cell libraries) were prepared in parallel, using the same conditions and reagents. Using a Bravo automated liquid handling platform (Agilent), mRNA was reverse transcribed into cDNA with an oligo containing unique molecular identifiers and cell barcodes. Uniquely barcoded cDNAs were pooled (each pool representing half of the original 16x 384-well plates, giving a total of 32 pooled batches in total) and linearly amplified using T7 *in vitro* transcription. Resulting RNA was fragmented and ligated to a second oligo containing a barcode specific to that pool and the Illumina sequences, using T4 RNA Ligase I. Finally, RNA was reverse transcribed back into DNA and PCR amplified with 17x cycles. For primer sequences see Jaitin *et al.* (2014)[88].

2.10.3 Sequencing

cDNA libraries were tested for efficient amplification, and fragment size distribution and concentration calculated using Qubit and Tapestation. Single cell RNA-seq libraries were pooled at equimolar concentration and

sequenced as a paired end run on an Illumina NextSeq 500 with a 75 cycle v2 kit. Read one was 59bp (covering the pool-specific barcode and cDNA), and read two 17bp (covering the well-specific barcode and UMI).

2.10.4 Pre-processing and filtering of MARS-seq reads

Reads were mapped against the *Xenoturbella bocki* genome using bowtie2 with default parameters and genes annotated based on Trinity assembly (Yann Loe Mie, unpublished data). After annotation, gene intervals were extended up to 4kb downstream or until the next in-frame gene was found. This was owing to poor 3' annotation of the *Xenoturbella* genes, which caused many of the MARS-Seq reads (which is a 3' biased method of RNA-Seq) to map outside of the annotated genes. Mapped reads were subsequently processed and filtered as is described in the MARS-Seq protocol.

Reads were deduplicated based on a unique molecular identifier (UMI) sequence to account for duplicates that might arise as a result of synthesis or sequencing errors. Only reads associated with a unique UMI tag were assigned a read count for a particular gene. Outlier cells were removed from downstream analysis based on the log UMI counts per cell vs. cell size. This eliminated low quality cells, which had UMI counts that did not correlate with the cell size.

2.10.5 Clustering single cell libraries with Seurat

Unsupervised clustering of the single cell libraries was carried out using Seurat[89], with the inclusion of cells that had a total UMI count of 100 or above. This retained 5006 cells for analysis in the clustering pipeline. Genes were only included for further analysis if they were identified in at least three of the retained cells. The Seurat package was run using recommended parameters to identify highly variable genes across the single cell libraries, based on 15 dimensions of variance in the data. tSNE non-linear dimensional

reduction of variance across the data set was carried out to represent this variance as a tSNE scatter plot of the single cell libraries.

2.11 Tomoseq Protocol

2.11.1 Cryosectioning of *Xenoturbella*

Animals were placed into RNALater as described in 2.2.2.

Prior to sectioning, all necessary equipment (molds, tools, cryostat components) was cleaned thoroughly with 100% EtOH. In addition, bench surfaces, molds, tweezers and a clean microscope slide were sprayed with RNaseZAP and rinsed with deionised water.

A small rectangular mold was filled halfway with OCT mounting medium, with care taken to remove any air bubbles, and placed at -20°C for ten minutes to solidify. The animals required for TomoSeq was removed from RNALater solution and placed on the clean microscope slide. The anterior and posterior ends of the animal were identified based on photographs and a description made before it was anaethetised. Using clean tweezers, the animal was very briefly dipped into OCT medium and then placed onto the frozen OCT block, on ice. The location and orientation of the animal was noted on the mold, and the entire animal covered in OCT, so that it was completely embedded in the mounting medium. The mold was placed at -20°C for half an hour to ensure it was completely frozen.

Once frozen, the OCT block was removed from the mold, and excess mounting medium cut away using a clean, sterile blade. The OCT block was then mounted on the cryostat with the anterior end closest to the blade. Cryostat sectioning was carried out at -18°C, cutting at sections of 15µM. Groups of an appropriate number of sections (see 2.11.2) were together placed directly into 1ml Trizol solution in Eppendorf LoBind tubes at RT using clean, autoclaved toothpicks, until the whole animal was sectioned along the

antero-posterior axis. Trizol tubes were stored at 4°C for one hour and vortexed briefly to ensure total dissolution of tissue sections before being stored at -80°C until required for RNA extraction.

2.11.2 RNA extraction

Working in batches of ten, tubes were removed from -80°C storage and allowed to reach room temperature over a period of approximately 1 hour. To each tube, 200µl of chloroform was added and the solution vortexed for ~20 secs until a homogenous opaque pink colour, then left standing at room temperature for three minutes. Tubes were centrifuged at 13,000 rpm at room temperature for ten minutes. RNA contained in the uppermost aqueous layer was removed into 1.5ml LoBind tubes (approximately 500µl), and an equal volume of freshly made 80% EtOH (diluted in nuclease-free water), added and gently mixed into the aqueous layer until no immiscible separation was visible.

Depending on the total number of sections transferred into Trizol, either the Qiagen RNeasy Mini kit (180µM total tissue) or Qiagen RNeasy Micro kit (60µM total tissue) was used for RNA extraction. The same protocol was followed for both kits, following manufacturer-recommended guidelines. Having troubleshooted the RNA-extraction protocol, the only protocol amendment was to place columns in clean collection tubes following centrifuge steps, instead of discarding the flow-through.

An appropriate volume of RNase Inhibitor (calculated as 2U/µl RNA) was added directly into the LoBind tube into which RNA was to be eluted. After the wash steps, columns were placed into the LoBind tube containing RNAse inhibitor, and nuclease-free water pipetted directly onto the membrane of the column (10µl for the Micro kit, and 30µl for the Mini kit). Columns were left to stand for one minute, and then centrifuged at room temperature at 13,000 rpm for one minute. Samples were quantified on a NanoDrop or a Qubit and stored at -80°C until required.

2.11.3 CelSeq2: linear amplification of RNA

2.11.3.1 Protocol establishment

Linear amplification of *Xenoturbella* RNA was implemented using the CelSeq2 protocol, with some modifications necessary for a larger starting input of RNA. To establish the protocol, an initial trial experiment of RNA extracted from 20 tissue sections of ~192μM (12x 16μM) was carried out. A series of practice CelSeq2 protocols were run using various parameters, including: different starting quantities of RNA; pooling different numbers of barcoded sections prior to IVT; and running various numbers of cDNA library preparation PCR cycles. These initial experiments showed that the protocol was best suited for initial RNA input of low concentrations: higher concentrations of starting RNA resulted in cDNA libraries that were broad and flat on the Bioanalyzer with no clear peaks (see Appendix 6). Consequently, all RNA samples were quantified using Qubit, and a final concentration of 0.4ng RNA per section used for subsequent amplification. Following mRNA amplification and cleanup, all 20 sections were pooled prior to IVT.

In order to choose the optimal number of PCR library preparation steps, I ran the pooled cDNA library preparation as test PCRs with amplification cycles increasing from x9 to x17 (see Appendix 6). Optimal library PCR amplification was determined based on selecting the PCR cycle number in the middle of the exponential amplification stage: too few cycles would result in under-amplification of the library; too many cycles would cause a loss of library complexity, and a large trailing peak on the Bioanalyzer cDNA library data curve (see Appendix 6). 11 and 13 cycles were chosen as subsequent practice runs for the final 20-section cDNA library. As a modification to the original CelSeq2 protocol, I used bead cleanup of the resulting libraries with a more stringent size-selection bead

ratio, so as to select for fragments of greater than 200 base pairs, and avoid noise in the library caused by primer dimers. Based on these practice parameters, the optimal library (13x PCR cycle amplification, stringent size-selection bead clean up) was sequenced at low coverage to verify successful library preparation and the presence of *Xenoturbella* transcripts

2.11.3.2 RNA preparation and amplification

Initial primer mixes or 6ul were prepared for each RNA sample, with starting RNA input scaled to be approximately equal to 2ng for each section:

Component	Volume
Primer (25ng/µl)*	1µl
ERCC Spike-in (Ambion)**	1µl
dNTPs (10mM)	0.5µl
Nuclease-free water	2.5µl
Clean RNA (~2ng)	1µl

*For primer sequences see supplementary information
**ERCC Spike-in was added at a dilution of 1:10,000

1.2µl of the above primer mix was taken for each section and incubated at 65°C for five minutes in a thermal cycler with the heated lid set to 65°C. Samples were spun down briefly and placed on ice, and 0.8µl of the following mix added to each reaction:

Component	Volume
First strand buffer (Invitrogen)	0.4µl
DTT 0.1M (Invitrogen)	0.2µl
RNaseOUT (Invitrogen)	0.1µl
Superscript II (Invitrogen)	0.1µl

Samples were incubated at 42°C for one hour in a thermal cycler with the lid set to 50°C. The enzyme was then heat inactivated by heating to 70°C for 10 minutes.

2.11.3.3 Second strand reaction

Tubes were removed from 70°C incubation and placed on ice. 10µl of the following mix was added to each reaction:

Component	Volume
DDW	7µl
Second strand buffer (Invitrogen)	2.31µl
dNTP (10mM)	0.23µl
E. coli DNA ligase (Invitrogen)	0.08µl
E. coli DNA polymerase (Invitrogen)	0.2µl
RNaseH (Invitrogen)	0.08µl

Tubes were spun down briefly and incubated at 16°C for two hours in a thermal cycler with an unheated lid.

2.11.3.4 Pooling and cDNA Cleanup

Samples to go to the same IVT were pooled together. 20 differently barcoded primer sequences were used (see Appendix 7). Samples were pooled in batches of 4x 20 and 1x 10 sections (see Figure 8.4) along the antero-posterior axis.

AMPure XP beads (Beckman Coulter) were warmed to room temperature and vortexed thoroughly to ensure dispersal of the beads. To each pooled sample of 20 sections, 1.2x volume of beads were added and pipetted gently to mix well. Samples were incubated at room temperature for ten minutes and then placed on a magnet for five minutes, until the liquid was clear and all of the beads were bound to the wall of the tube. The supernatant was removed without disturbing the beads, and the beads washed twice in 200µl of freshly prepared 80% EtOH. All of the final EtOH wash was removed from the beads and they were left to air dry for five minutes, until no liquid drops were visible in the tube. Beads were resuspended in 6.4µl nuclease free water, pipetted up and down to mix thoroughly, and incubated at room temperature for five minutes. The tubes were placed back on the magnetic stand until the liquid became clear, and the supernatant transferred to a clean tube.

2.11.3.5 IVT

9.6µl of the following mix was added to each tube (all reagents from the Ambion T7 transcription kit):

Component	Volume
ATP	1.6µl
GTP	1.6µl
CTP	1.6µl
UTP	1.6µl
10x T7 buffer	1.6µl
T7 enzyme	1.6µl

Samples were incubated for 13 hours at 37°C in a thermal cycler with the lid set to 70°C.

2.11.3.6 EXO-SAP treatment

6µl of EXO-SAP enzyme (Affymetrix) was added to each tube, and incubated at 37°C for 15 minutes to remove excess primers from the sample.

2.11.3.7 RNA Fragmentation

5.5µl of fragmentation buffer was added to each sample. Tubes were incubated at 94°C for three minutes and then placed immediately on ice. 2.75µl 0.5M EDTA pH8 was added to each sample to stop the fragmentation.

2.11.3.8 aRNA Cleanup

RNAClean XP beads (Beckman Coulter) were brought to room temperature and vortexed thoroughly. 1.8x volume beads were added to each aRNA sample and the samples incubated at room temperature for 10 minutes. Bead cleanup was carried out as described for cDNA, but with three repeats of a 70% EtOH wash. Beads were resuspended in 7µl nuclease free water and the final supernatant transferred to a new tube for library preparation.

2.11.4 Library preparation

2.11.4.1 RT Reaction

5µl of clean aRNA was transferred to a new tube. 1µl randomhexRT primer and 0.5µl dNTPs (10mM) were added to each sample. Tubes were incubated for five minutes at 65°C in a thermal cycler with the heated lid set to 65°C. Samples were immediately placed on ice, and 4µl of the following mixture added to each tube:

Component	Volume
First strand buffer	2µl
DTT 0.1M	1µl
RNaseOUT	0.5µl
Superscript II	0.5µl

Samples were incubated for ten minutes in a thermal cycler at 25°C with the heated lid turned off, followed by one hour incubation at 42°C in a thermal cycler with the heated lid set to 50°C.

2.11.4.2 PCR amplification

For each sample, half of the RT reaction was used for PCR amplification. Pooled barcoded samples were amplified using Illumina primer RP1, and separate Illumina RPIX primer sequences for each of the five libraries chosen as directed in the Illumina pooling guide (see Appendix 7). PCR reagents (Phusion High Fidelity PCR Kit, New England Biolabs) were prepared as follows:

Component	Volume
Polymerase	0.25µl
HF Buffer	5µl
dNTPs	0.5µl
RT Template	5µl
RP1	1µl
RPIX	1µl
N-free H_2O	12.25µl

Samples were amplified in a thermal cycler using the following PCR cycling protocol: 30 seconds at 98°C; X number* of cycles of 10 seconds at 98°C, 30 seconds at 60°C and 30 seconds at 72°C; 10 minutes at 72°C; and a final hold stage at 4°C.

***** The number of PCR cycles for each library was determined as outlined in section 2.11.3.1.

2.11.4.3 Library Cleanup

AMPure XP beads were brought to room temperature and vortexed thoroughly. 0.9x volume of beads was added to the PCR product and incubated at room temperature for 15 minutes. Bead cleanup was carried out as described for cDNA. Beads were eluted in 25µl nuclease free water, and the supernatant transferred to a new tube. Bead cleanup was repeated with 0.9x volume beads, resuspending in 10µl nuclease free water. The final 10µl library was transferred to a new tube and tested for efficient amplification, fragment size distribution and concentration using Qubit and Bioanalyzer. All five libraries, identifiable by different RPIX primers, were pooled at equimolar concentration for sequencing.

2.11.5 Sequencing and mapping

Computational analysis of data from both rounds of Tomoseq was carried out in conjunction with Philipp Schiffer (Telford lab).

Initial RNA samples which I did not linearly amplified using the CelSeq2 protocol were prepared for sequencing remotely using the SmartSeq2[90] protocol. For all cDNA library sequencing, 100bp paired-end sequencing was carried out on an Illumina HiSeq machine.

2.11.5.1 Mapping and analysis of SmartSeq2 reads

Paired end reads were assembled using CLC Assembly Cell (http://www.clcbio.com/products/clc-assembly-cell/) to create a reference transcriptome. Read counts were normalised to transcripts-per-million (tpm)

using kallisto[91] and differential expression analysis for the genes-of-interest implemented using sleuth[92].

2.11.5.2 Computational analysis of CelSeq2 Tomoseq data

Forward reads from all five libraries were sorted by their 6bp barcode sequence using a custom Julia script (written by Philipp Schiffer, Telford lab) and the corresponding reverse reads extracted using pullseq (https://github.com/bcthomas/pullseq). Reads were normalised and mapped to the genome using kallisto, as in the initial analysis, and mapped reads deduplicated based on their UMI sequence using UMI_TOOLS (https://github.com/CGATOxford/UMI-tools). Normalisation and expression pattern analyses were conducted using StringTie (https://ccb.jhu.edu/software/stringtie/) and Ballgown (https://github.com/alyssafrazee/ballgown).

3 Acoelomorpha mitochondrial genomes

The results of this chapter are published as: Robertson *et al.* The mitochondrial genomes of the acoelomorph worms *Paratomella rubra*, *Isodiametra pulchra* and *Archaphanostoma ylvae*. *Scientific Reports* (2017)[93]. (Appendix 9). Permission to reproduce this has been granted by *Scientific Reports*.

In addition, as a separate project I sequenced the mitochondrial genome of the geophilomorph centipede *Strigamia maritima*: Robertson *et al*. The complete mitochondrial genome of the geophilomorph centipede *Strigamia maritima*. *PLOS ONE,* 10 (2015)[94]. (Appendix 9). Permission to reproduce this has been granted by *PLOS ONE*. This work also contributed to the *Strigamia maritima* nuclear genome paper[95].

3.1 Introduction

3.1.1 Molecular data in the phylogenetic inference of the Xenacoelomorpha

As described in 1.2.3, two hypotheses prevail regarding placement of the Xenacoelomorpha in the Bilateria. A number of different lines of molecular data have been used to infer their phylogenetic position, including microRNAs, transcriptomic data, ESTs, and mitochondrial genes[21,28,30,96]. Data from the Xenacoelomorpha are prone to systematic and stochastic error in phylogenetic reconstruction[3,28]. Of all sources of systematic error, long branch attraction (LBA) is the most well known and also the most pervasive[28]. Long branch attraction occurs when one lineage has a much faster rate of sequence evolution than the others. This results in a greater number of character changes, and causes the 'branch' of that lineage to

appear longer. The 'long branch' taxa may therefore group with a second 'long branch' fast-evolving lineage, or to a long-branched, but not necessarily fast-evolving, distant outgroup. It is well known that the Acoelomorpha are a rapidly evolving taxon[28]. Consequently, the basal position that is sometimes resolved for Xenacoelomorpha could be an artefact resulting from the attraction between this long-branched lineage and a long-branched outgroup, causing them to be dragged down to the base of the tree (Figure 1.6). Whilst systematic error is largely caused by a failure of the phylogenetic model to properly account for the characteristics of the data, one way to reduce stochastic error is to increase the size of the data set.

Mitochondrial DNA has a number of features that make it a valuable tool for phylogenetic inference. Its small size and consistent gene complement make it relatively affordable and quick to sequence and annotate in comparison to nuclear genomes, and this is a particular advantage for phyla that lack a wealth of comprehensive nuclear genomic data – including the Xenacoelomorpha. The recent discovery of four new *Xenoturbella* species has contributed four further complete mitochondrial genomes[23], but available mitochondrial data for the Acoela remains very poor, with just one complete genome of *S. roscoffensis* representing the group[19].

3.1.2 Mitochondrial genomes

Metazoan mitochondrial genomes are closed-circular molecules typically comprising 37 genes which are, for the most part, invariant across the Metazoa[28]. These comprise two rRNAs of the mitochondrial ribosome; 22 tRNAs necessary for translation; and 13 protein-coding genes for the enzymes of oxidative phosphorylation (*cytochrome oxidase c subunit* (*cox*) *1, 2, 3*; *NAD dehydrogenase* (*nad*) *subunit 1, 2, 3, 4, 4l, 5, 6*; *cytochrome b* (*cob*); *atp8* and *atp6*). *atp8* is the only gene known to have been commonly lost from this complement, and this has been observed in a number of independent metazoan lineages, including the acoel *S. roscoffensis*[19].

3.1.3 Role of mitochondria in phylogenetic inference

Concatenated mitochondrial protein-coding gene sequences have been used extensively for phylogenetic inference in recent years, and this has proved beneficial in resolving a number of contested evolutionary relationships. Most recently, Rouse *et al.* (2016)[23] used mitochondrial protein-coding sequence data from four newly discovered species of *Xenoturbella* (*X. hollandorum*, *X. churro*, *X. monstrosa*, and *X. profunda*) to infer the internal phylogeny of the Xenoturbellida. Furthermore, wider phylogenetic inference including mitochondrial proteins from these species placed Xenacoelomorpha with the deuterostomes, corroborating previous mitochondrial phylogenetic analysis of this phylum[23,28,96]

In addition to protein-coding gene sequence data, other features of mitochondrial genomes can be used for reconstructing phylogenetic relationships. These include variation in mitochondrial genetic code[97]; a higher rate of sequence evolution than nuclear genomes[98]; and changes to the secondary structure of rRNAs and tRNAs[99]. Comparing the arrangement of genes within the mitochondrial genome of different species can also be a powerful tool in the analysis of phylogenetic relationships[99]. As described, mitochondrial gene content is largely invariant across the Metazoa. The order in which these genes are arranged is fairly stable, and have been conserved for up to hundreds of millions of years in some metazoan lineages. Rearrangement events, thought to occur via a model of 'duplication and deletion', whereby a portion of the mitochondrial genome is duplicated, and the original copy of the duplicated gene subsequently deleted, are rare[99,100]. The infrequency of such rearrangements, and the huge number of possible rearrangement scenarios, means that convergence on the same gene order in unrelated lineages is unlikely. Gene order is thus likely to retain evolutionary signals, with a common gene order being indicative of common ancestry and informative for the study of metazoan divergence[101]. Rearrangement of genes within the mitochondrial genome of different species can be a particularly powerful tool in the analysis of phylogenetic relationships, and may also indicate accelerated evolution in a taxon.

Despite these advantages, phylogenies based on mitochondrial protein-coding gene sequences are prone to processes – such as compositional heterogeneity and accelerated substitution rates - that may lead to tree reconstruction artefacts and misleading phylogenetic signals if not properly accounted for[102]. Compositional heterogeneity in mtDNA arises most commonly as a result of the mtDNA repair system, which is inefficient at replacing adenine nucleotide insertions, resulting in a reduction of G/C nucleotides in comparison to A/T. Further heterogeneity may occur as a result of the asymmetrical replication mechanism of mtDNA[103]. During replication the lagging 'heavy' (H-) strand is left unpaired for a long time, leaving it vulnerable to deamination (A to G and C to T). This increases the proportion of G and T nucleotides present in the lagging strand, leaving the 'light' (L-) strand with a greater number of A and C nucleotides, with the difference between the two described as GC- and AT- skew[104]. Compositional heterogeneity and GC- AT- skew should therefore be considered as potential limitations of direct mitochondrial sequence comparisons[105,106]. Accelerated substitution rates in mitochondrial DNA can also lead to misinformation owing to the clustering of rapidly evolving lineages (LBA)[107]. This is of particular relevance for phylogenetic inference of the Acoela, as acoel species demonstrate a very rapid rate of nucleotide substitution compared to other metazoans, leaving them vulnerable to LBA and incorrect clustering in a basal position[15,108]. 'Breaking' long branches is therefore important to minimise reconstruction artefacts. This can be achieved by: the broad sampling of taxa; selecting 'short branch' representative species; using models designed to overcome LBA artefact; and careful selection of outgroup species.

3.1.4 Aims of Chapter

The aim of this chapter was to sequence the mitochondrial genomes from three species of Acoela: *P. rubra*, *I. pulchra*, and *Archaphanostoma ylvae* (Figure 1.3). Adult specimens of all animals are approximately 1mm in length, and, as is typical for small acoel species, they occupy the littoral and

sub-littoral zones of marine ecosystems: *P. rubra* has been described across Europe and North America[109,110]; *I. pulchra* lives abundantly in the mud flats of Maine but has been maintained in long-term culture for a number of years[70]; and *A. ylvae* has been described from the West coast of Sweden[111]. All species move freely within the sediment by gliding on a multiciliated epidermis.

First described by Rieger and Ott (1971) *P. rubra* is an elongate and flattened worm belonging to the family Paratomellidae[109,110]. A 9.7kb fragment of mitochondrial genome has previously been described from specimens of *P. rubra* collected on the Mediterranean coast of Spain[20]. *I. pulchra* belongs to the family Isodiametridae: no mitochondrial sequence data has been published for *I. pulchra*, but its role as a putative 'model acoel' makes this species particularly valuable for investigation[72]. *A. ylvae* also belongs to the family Isodiametridae family of acoels. Originally described by Kånneby *et al.* (2014), this species has not been the subject of extensive investigation: its *cox1* gene has been sequenced and used for classification, but no further genes from its mitochondrial genome have been sequenced[111].

Sequencing the mitochondrial genomes of these three taxa should make a significant contribution to the amount of mitochondrial molecular data available for the Acoela. In addition, *P. rubra* represents a comparatively slowly evolving acoel species, making it less prone to LBA than other fast-evolving acoel species. I aimed to analyse mitochondrial gene content and gene order in these three species in comparison to other metazoan species, and use protein-coding gene sequences to carry out phylogenetic analysis with sequence alignments from other mitochondrial protein-coding genes.

3.2 Results

3.2.1 Genomic Composition

3.2.1.1 Paratomella rubra

I was successful in assembling 14,954 base pairs of the *P. rubra* mitochondrial genome, from an initial starting point of three genome assembly fragments, and using Sanger sequencing of subsequent PCR fragments (see section 2.4 and Appendix 1). Although I was unable to close the circular genome of *P. rubra*, the final sequence contains all 13 protein-coding genes, both ribosomal genes, and 22 putative tRNAs. Compared to the fragment of genome previously published, this analysis contributes four addition protein-coding genes and 12 addition tRNAs[20]. All genes are found exclusively on one strand of the sequence. Allowing for overlap between genes, protein-coding genes account for 74.79% of the genomic sequence; ribosomal genes 13.95%; tRNAs 9.10%, and non-coding DNA just 2.04%. One 'long' non-coding stretch of 156 nucleotides is found between *cox2* and *nad1* (Figure 3.1A; Table 1).

Table 1: Organisation of the *P. rubra* 14.9kb mitochondrial genome sequence. All genes found on the 'plus' strand.

Feature	Strand	Start	Stop	Length (bp)	Length (AA)	Start Codon	Stop Codon	Intergenic region
trnH (gtg)	+	368	426	59				7
trnP (tgg)	+	434	495	62				0
cox1	+	496	2058	1563	521	ATA	TAA	5
trnT (tgt)	+	2064	2123	60				7
nad2	+	2131	3105	975	325	ATT	TAG	-4
nad6	+	3102	3563	462	154	ATA	TAA	-101
rrnL	+	3463	4819	1357				-53
trnL1 (tag)	+	4767	4824	58				4
trnG (tcc)	+	4829	4887	59				0
atp6	+	4888	5496	609	203	ATA	TAG	-10
trnV (tac)	+	5487	5553	67				0
atp8	+	5554	5730	177	59	ATT	TA-	9
cox2	+	5740	6402	663	221	ATT	TAA	156
nad1	+	6559	7602	1044	348	ATT	TAA	-104
trnS2 (tga)	+	7499	7568	70				32
trnD (gtc)	+	7601	7662	62				2
trnI (gat)	+	7665	7728	64				0
trnN (gtt)	+	7729	7798	70				-2
trnF (gaa)	+	7797	7856	60				-39
rrnS	+	7818	8547	730				-6
trnR (tcg)	+	8542	8608	67				0
trnM (cat)	+	8609	8669	61				68
cox3	+	8738	9523	786	262	ATT	TAA	3
trnY (gta)	+	9527	9585	59				0
cob	+	9586	10668	1083	361	ATA	TAA	8
trnL2 (taa)	+	10677	10737	61				-2
trnS1 (gct)	+	10736	10799	64				-6
nad4	+	10794	12119	1326	442	ATC	TAA	4
trnA (tgc)	+	12124	12181	58				10
nad3	+	12192	12551	360	120	ATT	TAG	18
trnC (gca)	+	12570	12629	60				6
nad5	+	12636	14387	1752	584	ATA	TAG	6
trnQ (ttg)	+	14394	14449	56				1
trnE (ttc)	+	14451	14510	60				0
trnK (ttt)		14511	14576	66				-18
nad4l	+	14559	14867	309	103	ATA	TAA	2
trnW (tca)	+	14870	14933	64				

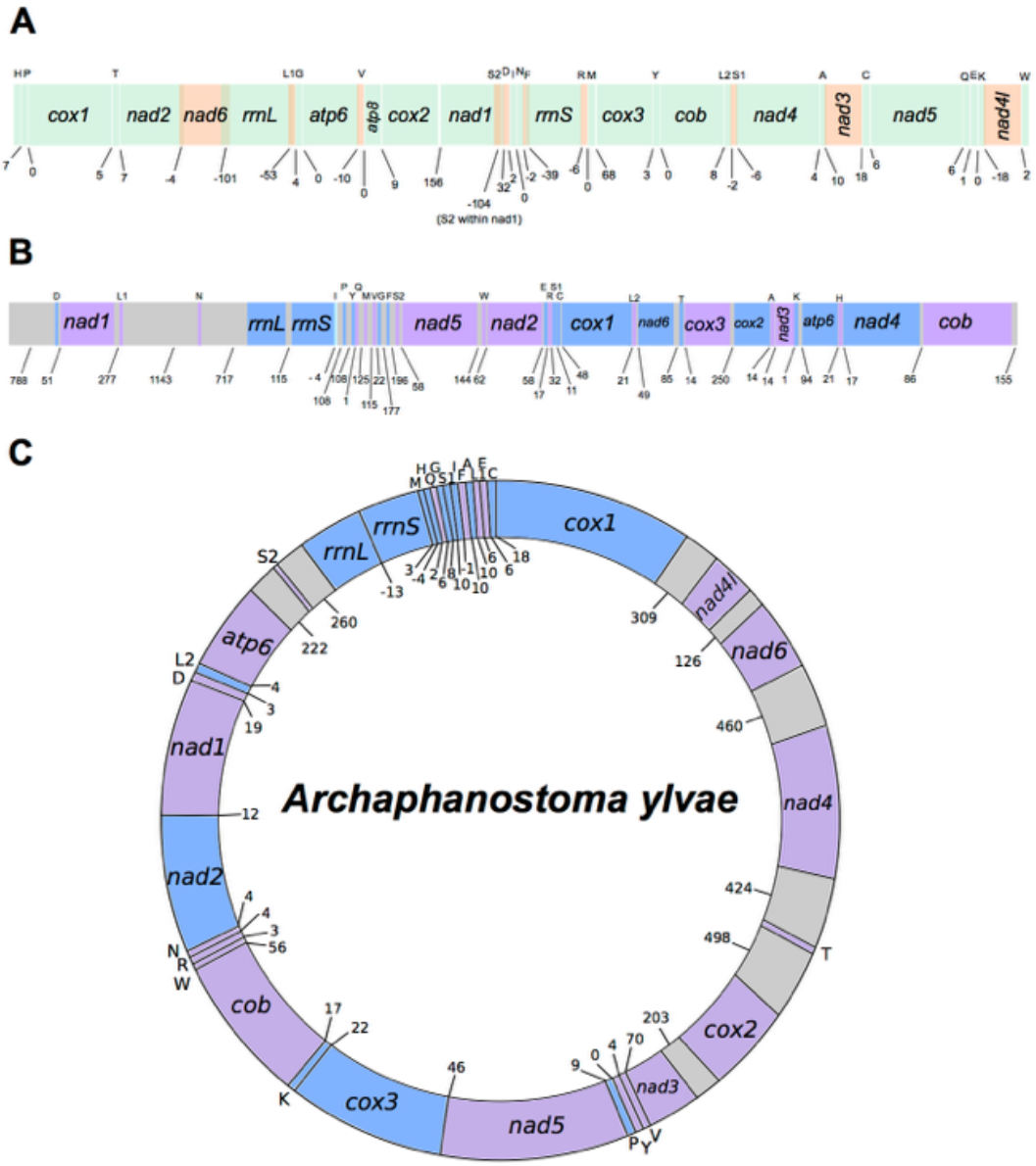


Figure 3.1. Overview of the mitochondrial genome sequences resolved for *P. rubra*, *I. pulchra* and *A. ylvae*. Genes not drawn to scale. Numbers beneath the sequence show intergenic spaces (positive values) or intergenic overlap (negative values). Protein-coding genes are denoted by three-letter abbreviations; ribosomal genes by four-letter abbreviations. tRNAs are shown by single uppercase letters. (A) *P. rubra* 14,957 bp sequence. All genes found on the 'plus' (forward) strand. Where genes, rRNAs or tRNAs are shown in orange, this is solely to demonstrate overlap with the adjacent genes, rRNAs or tRNAs. (B) *I. pulchra* 18,725 bp consensus sequence. Genes found on the 'plus' (forward) strand are shown in blue; genes on the 'minus' (reverse) strand are shown in purple. (C) *A. ylvae* 16,619 bp mitochondrial genome. Genes found on the 'plus' (forward) strand are shown in blue; genes on the 'minus' (reverse) strand are shown in purple. Non-coding regions greater than 100 nucleotides in length are shown in grey.

In *P. rubra*, *trnS2* is predicted entirely within the sequence coding for *nad1*, and the secondary structure of this sequence has deviated from the traditional 'cloverleaf' shape that is expected for tRNA. In addition, three of the other predicted tRNA sequences have minor overlaps with protein-coding genes: *trnA* with *nad3* (20 nucleotides); *trnK* with *nad4l* (18 nucleotides); and *trnS1* with *nad4* (six nucleotides). All but five nucleotides of *trnL1* are predicted within the same sequence as *rrnL*. With the exception of *trnT*, all predicted tRNAs have an amino-acyl acceptor stem composed of seven base pairs, and all predicted tRNAs apart from *trnT* and *trnS2* have a five base pair anticodon stem (*A,C,G,I,K,L1,L2,P,Q,R,T*). All tRNAs have a DHU arm of three or four nucleotides. The structure of the TΨC arm shows greater variability, with a number of tRNAs having either a truncated stem, or the arm entirely lacking (Figure 3.2).

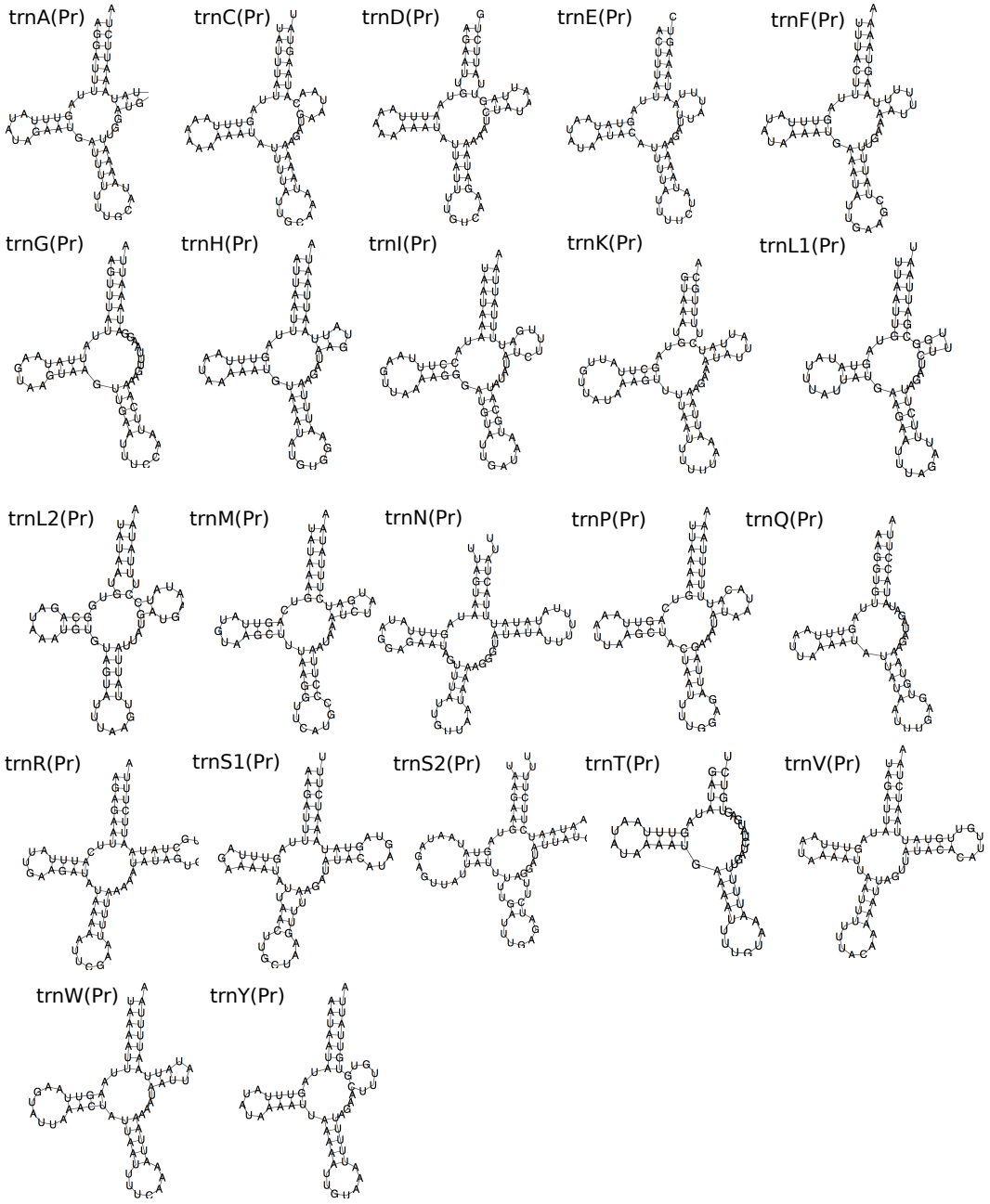


Figure 3.2. Predicted secondary structure of tRNAs from the mitochondrial genome sequence of *P. rubra.* Structures predicted by MiTFi in Mitos.

3.2.1.2 Isodiametra pulchra

As an initial starting point for the *I. pulchra* genome, three contigs of mitochondrial sequence of lengths 13kb, 3.5kb, and 19kb, were recovered from transcriptome sequencing data. After aligning these sequences, I found that the entire 13kb contig, and 2.4kb of the 3.5kb contig were perfectly matching subsets of the longer 19kb sequence (Figure 3.3). I designed several sets of PCR primers to try and verify the sequence between the 3' end of the 13kb, and 5' end of the 3.5kb fragments also found on the long 19kb sequence. However, despite trying numerous PCR primer combinations and different PCR protocols, no PCR amplification successfully bridged the sequence between the 13kb and 3.5kb fragment. I found that the last (3') 300bp of the 13kb fragment was duplicated in the opposite orientation within the end (3') region of the 3.5kb fragment. Although the long 19kb fragment contained the repeated region between the 13kb and 3.5kb fragments, no PCR amplification strategy was successful in connecting sequences flanking the repeated region (Figure 3.3). Any Sanger sequencing fragments that did partially cover the repeated regions were ambiguous. Consequently, I focused instead on verifying the sequence of the two shorter 3.5kb and 13kb contigs and on amplifying and sequencing the region lying between them, instead of using the 19kb sequence as a 'template'. In doing this, I reconfirmed the majority of the 13kb fragment using PCR amplification and Sanger sequencing. I also amplified and sequenced fragments joining the 3' end of the 3.5kb fragment with the 5' end of a 1.3kb fragment containing the *rrnL* gene, which was identified using a BLAST query for this gene in the transcriptomic sequence data (Figure 3.3).

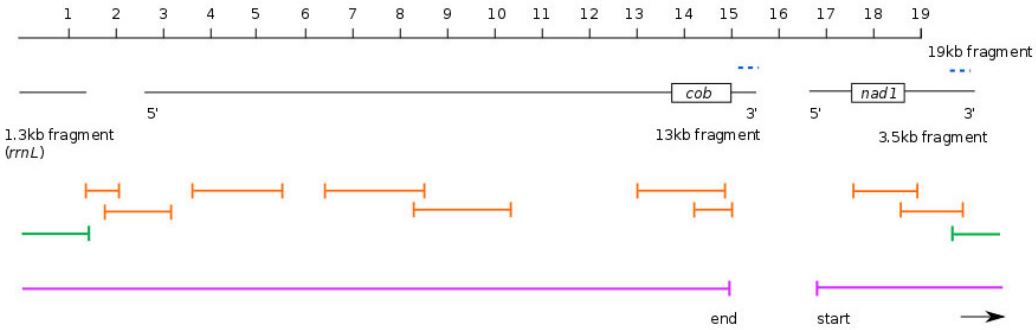


Figure 3.3 Overview of the initial transcriptome assembly fragments and PCR strategy for scaffolding the *I. pulchra* mitochondrial genome. 1.3kb, 13kb and 3.5kb fragments aligned to a continuous 19kb fragment, with the location of the duplicated sequence in the 13kb and 3.5kb fragments shown by blue dashed lines. The 'start' and 'end' regions of the 13kb and 3.5kb fragments are annotated by 5' (start) and 3' (end). The approximate location of cob and nad1 protein-coding sequence are shown for reference. Reliable PCR-amplicons are shown in orange; the green PCR fragment indicates successful joining of the 3' end of the 3.5kb fragment to the rrnL fragment, including the duplicated section. The 18,725 base-pair long sequence we resolve is indicated by the pink lines, from 'start' to 'end'.

From this re-focused analysis, I recovered the *I. pulchra* mitochondrial genome to be a minimal length of 18,725 base pairs, based on the transcriptomic data that could be validated by successful PCR amplification. This covers the region from the start of the 5' end of the 3.5kb sequence, linked through PCR fragments to the 5' end of the 13kb sequence, and up to the start of the duplicated sequence at the 3' end of the 13kb sequence (Figure 3.1). The region between the 3' end of the 13kb fragment and the 5' end of the 3.5kb fragment could not be spanned continuously by any PCR amplification, and so I was unable to confirm the validity of the duplicated sequences at this position, or completely close the circle of the mitochondrial genome. Therefore it is likely that the complete mitochondrial genome of *I. pulchra* is larger than the ~19kb sequence that could be verified by PCR, and may also include the duplicated sequence. Nonetheless, the verified 18,735 base pair sequence contains both ribosomal genes, all tRNAs and 11 protein-coding genes, found on both the plus and minus strands. No sequences resembling either *atp8* or *nad4l* could be found in this sequence (Figure 3.1B, Table 2).

Within the 18.7kb sequence, protein-coding genes account for 56.66%; ribosomal genes contribute 8.15% and tRNA genes 7.77%. Compared to the *P. rubra* and *S. roscoffensis* mitochondrial genomes, intergenic space in the *I. pulchra* sequence is unusually high: non-coding DNA accounts for 22.27% of the sequence, including 14 intergenic regions that are longer than 100 base pairs (Table 2). Predicted sequences for *rrnS* and *trnI* overlap by four base pairs, but no other overlap was found between any tRNAs or with any other protein-coding genes. All predicted tRNAs have an amino-acyl acceptor stem composed of seven base pairs and a five base pair anticodon stem, with the exception of *trnE*, *trnF* and *trnS2*, which have an anticodon stem composed of only four base pairs. The structure of the DHU and TΨC show greater variability, and are composed of either three or four, or between three and six, base pairs respectively, across the 22 tRNAs. Whilst the TΨC arm is missing entirely in *trnQ*, and very truncated in *trnE*, *trnF*, *trnG* and *trnP*, more of the predicted tRNAs fit the stereotypical 'cloverleaf' secondary structure than has been found for other acoel species, including *S. roscoffensis* and *P. rubra* (Figure 3.4).

Table 2: Organisation of the *I. pulchra* 18.7kb mitochondrial genome. Only the sequence that could be verified by sequencing results is included for analysis.

Feature	Strand	Start	Stop	Length (bp)	Length (AA)	Start Codon	Stop Codon	Intergenic region
trnD (gtc)	+	789	848	60				51
nad1	-	900	1784	885	295	ATG	TAA	277
trnL1 (tag)	-	2062	2130	69				1143
trnN (gtt)	-	3274	3339	66				717
rrnL	+	4057	4657	601				115
rrnS	+	4773	5698	926				-4
trnI (gat)	+	5695	5768	74				108
trnP (tgg)	+	5877	5939	63				108
trnY (gta)	+	6048	6111	64				1
trnQ (ttg)	-	6113	6173	61				125
trnM (cat)	-	6299	6360	62				115
trnV (tac)	-	6476	6543	68				22
trnG (tcc)	+	6566	6627	62				177
trnF (gaa)	+	6805	6873	69				196
trnS2 (tga)	-	7070	7139	70				58
nad5	-	7198	8907	1710	570	ATG	TAA	144
trnW (tca)	-	9052	9118	67				62
nad2	-	9181	10233	1053	351	ATG	TAA	58
trnE (ttc)	+	10292	10355	64				17
trnR (tcg)	-	10373	10439	67				32
trnS1 (tct)	+	10472	10539	68				11
trnC (gca)	+	10551	10613	63				48
cox1	+	10662	12197	1536	512	ATA	TAG	21
trnL2 (taa)	-	12219	12286	68				49
nad6	+	12336	12811	476	159	ATG	T--	85
trnT (tgt)	+	12897	12963	67				14
cox3	-	12978	13775	798	266	ATG	TAA	250
cox2	+	14026	14640	615	205	ATA	TAA	14
trnA (tgc)	-	14655	14718	64				14
nad3	-	14733	15110	378	126	ATG	TAA	1
trnK (ttt)	+	15112	15178	67				94
atp6	+	15273	15955	683	228	ATA	TA-	21
trnH (gtg)	-	15977	16042	66				17
nad4	+	16060	17403	1344	448	ATG	TAA	86
cob	-	17490	18570	1081	361	ATA	T--	155

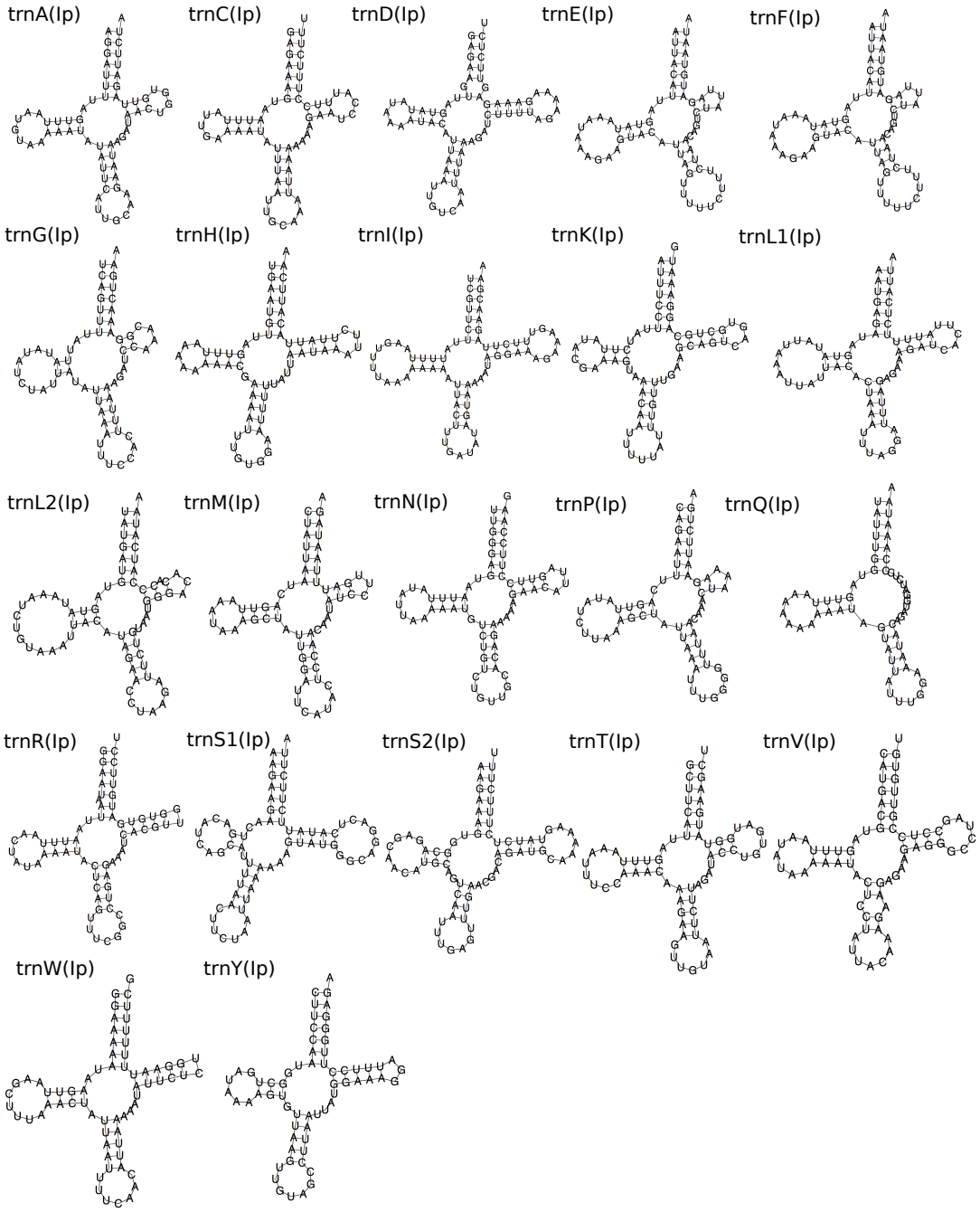


Figure 3.4. Predicted secondary structure of tRNAs from the mitochondrial genome sequence of *I. pulchra*. Structures predicted by MiTFi in Mitos.

3.2.1.3 Archaphanostoma ylvae

The complete closed circular mitochondrial genome of *A. ylvae* was recovered from genome sequencing data of *P. rubra* specimens collected from Yorkshire, UK. Contamination of the *P. rubra* samples was confirmed using NCBI BLAST, which found a 99% similarity to the sequence published for *A. ylvae cox1*. The complete *A. ylvae* mitochondrial genome is 16,619 nucleotides in length, containing 12 protein-coding genes, both rRNAs, and 22 predicted tRNAs (Figure 3.1C, Table 3). As is convention, with *cox1* at the start of the genome on the 'plus' strand, all other protein-coding genes apart from *cox3* and *nad2* are found on the 'minus' strand (Figure 3.1C). Both rRNAs are found on the plus strand, and tRNAs are distributed between the two. Accounting for a small amount of overlap between genes – found to total just 18 nucleotides across the whole genome – protein-coding genes make up 64.72% of the genome. tRNAs contribute 8.91%, and rRNAs 9.31%. As found for *I. pulchra,* non-coding DNA makes up a large amount of the genome, totalling 17.17%.

I identified putative sequences for all 22 mitochondrial tRNAs in the *A. ylvae* genome, although four of these (*trnE*, *trnI*, *trnK* and *trnS1*) have an e-value prediction of greater than 0.0001, and are therefore treated with caution. All predicted secondary structures of the tRNAs in the *A. ylvae* mitochondrial genome have standard-length acceptor and anticodon stems, and the majority – with the exception of *trnK*, *L1*, *L2*, *N*, *S2* and *Y* – have a four nucleotide long D-loop (Figure 3.5). As found in the mitochondrial genomes of *P. rubra* and *I. pulchra*, the greatest variability in structure is found in the TΨC arm, which is truncated in *trnD*, *E*, *F*, *L1*, *P*, *V*, *W* and *Y*, and missing entirely in *trnQ* (Figure 3.5).

Table 3: Organisation of the *A. ylvae* 16.6kb mitochondrial genome

Feature	Strand	Start	Stop	Length (bp)	Length (AA)	Start Codon	Stop Codon	Intergenic
cox1	+	1	1539	1539	513	ATA	TAA	309
nad4l	-	1849	2133	285	95	ATG	TAA	126
nad6	-	2260	2727	468	156	ATG	TAG	460
nad4	-	3188	4531	1344	448	ATA	TAA	424
trnT (tgt)	-	4956	5025	70				498
cox2	-	5524	6171	648	216	ATA	TAA	203
nad3	-	6375	6743	369	123	ATG	TAA	70
trnV (tac)	-	6814	6882	69				4
trnY (gta)	-	6887	6953	67				0
trnP (tgg)	+	6954	7022	69				9
nad5	-	7032	8687	1656	552	ATG	TAA	46
cox3	+	8734	9519	786	262	ATG	TAA	22
trnK (ttt)	+	9542	9611	70				17
cob	-	9629	10714	1086	362	ATT	TAA	56
trnW(tca)	-	10771	10837	67				3
trnR (tcg)	-	10841	10908	68				4
trnN (gtt)	-	10913	10984	72				4
nad2	+	10989	11990	1002	334	ATG	TAA	12
nad1	-	12003	12872	870	290	ATG	TAA	19
trnD (gtc)	-	12892	12954	63				3
trnL2 (taa)	+	12958	13025	68				4
atp6	-	13030	13731	702	234	ATA	TAA	222
trnS2 (tga)	-	13954	14020	67				260
rrnL	+	14281	15089	809				-13
rrnS	+	15077	15814	738				3
trnM (cat)	+	15818	15879	62				-4
trnH (gtg)	+	15876	15945	70				2
trnQ (ttg)	-	15948	16007	60				6
trnG (tcc)	+	16014	16079	66				8
trnS1 (tct)	+	16088	16151	64				10
trnI (gat)	+	16162	16231	70				-1
trnF (gaa)	-	16231	16296	66				10
trnA (tgc)	+	16307	16373	67				10
trnL1 (tag)	-	16384	16450	67				6
trnE (ttc)	-	16457	16522	66				6
trnC (gca)	+	16529	16601	73				18

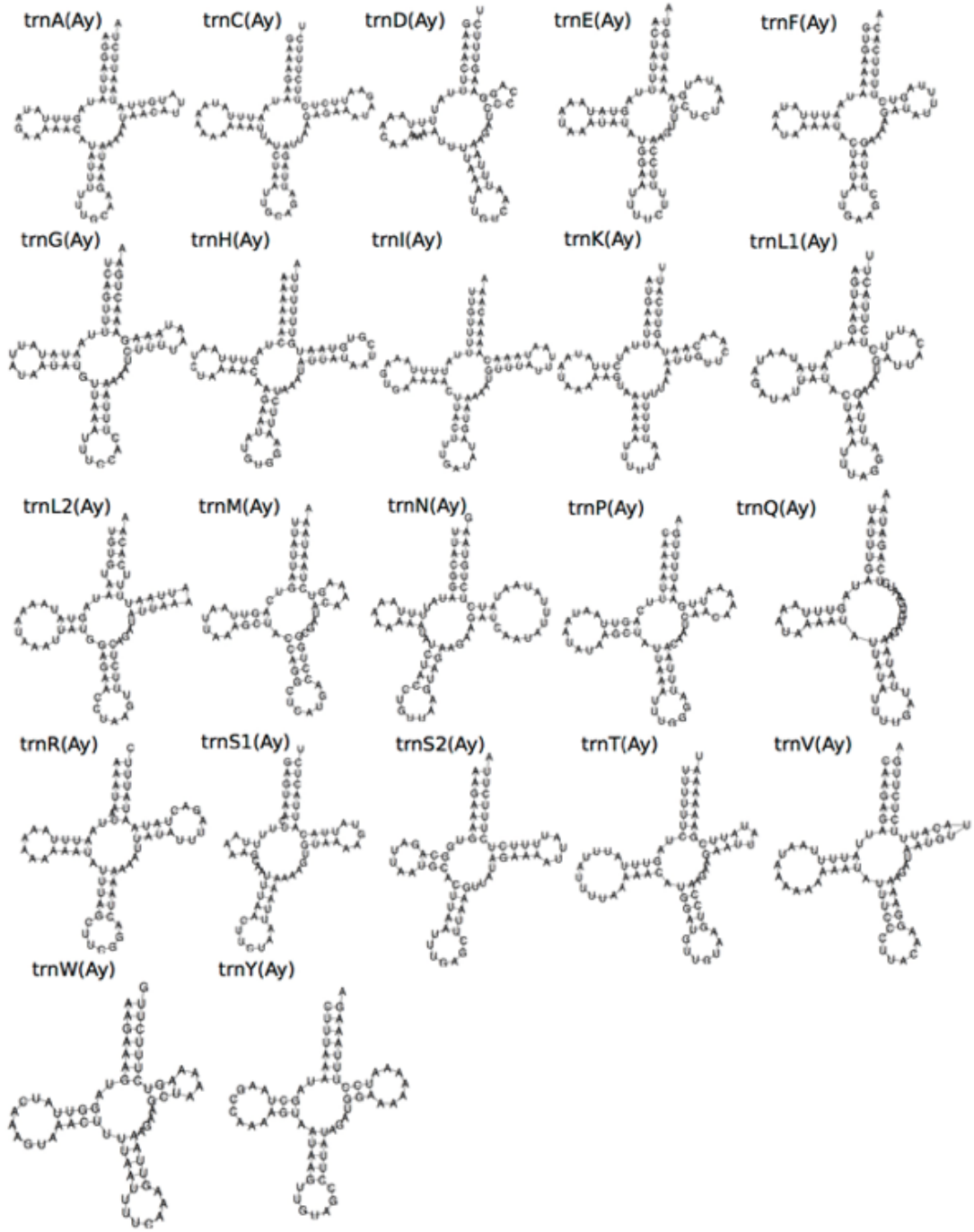


Figure 3.5. Predicted secondary structure of tRNAs from the mitochondrial genome sequence of *A. ylvae*. Structures predicted by MiTFi in Mitos.

3.2.1.4 Nucleotide composition of Acoelomorpha mitochondrial genomes

The *P. rubra* genome is 78.15% A+T rich, with is higher than the A+T content calculated for the *I. pulchra* genomic sequence at 67.28%, and the *A. ylvae* complete genome at 74.70%. Overall nucleotide usage on the plus strand of *P. rubra* (containing all coding sequence) is A = 29.29%, T = 48.86%, C = 6.77%, and G = 15.10%; GC-skew = 0.38 and absolute AT-skew = 0.25. Overall nucleotide usage for *I. pulchra* is: A = 34.04%, T = 33.24%, C = 16.45% and G = 16.27%; GC-skew = 0.006 and AT-skew = 0.012. For *A. ylvae*, A = 40.41%, T = 34.29%, C = 12.82% and G = 12.47%; GC-skew = 0.014 and AT-skew = 0.082. GC-skew and AT-skew absolute values for *P. rubra* are much higher than that of *S. roscoffensis*, although the absolute values for *I. pulchra* and *A. ylvae* are comparatively low[19]. AT-skew value for the *P. rubra* sequence is just 0.01 different from that of the published *P. rubra* mitochondrial genome fragment; GC-skew is slightly higher (published *P. rubra* GC-skew = 0.32)[20].

3.2.2 Gene order and gene arrangement

All thirteen protein-coding genes in *P. rubra* have complete initiation codons: ATA (x5) and ATT (x8) (Table 1). Five of the protein-coding genes that were previously published differ in the nucleotide sequence of their start codons: *nad2*, *atp8*, *cox2* and *cox3* all have ATA as an initiation codon in our analysis, compared to ATT found previously[20]. Twelve of the genes have full stop codons: TAA (x9) or TAG (x3) (Table 1). *atp8* was found to have a truncated stop codon (TA-), which is assumed to be completed during post-transcriptional modification. The eleven protein-coding genes found for *I. pulchra* also have full initiation codons: ATA (x4) and ATG (x7) (Table 2). Eight of the genes for this species have full stop codons: TAA (x7) and TAG (x1); *nad6*, *atp6* and *cob* are inferred to have truncated stop codons (Table 2). Initiation codons in *A. ylvae* are: ATA (x4), ATG (x7) and ATT (x1); all genes have TAA as stop codons, with the exception of *nad6*, which has TAG (Table 3). As in other invertebrate mitochondrial genomes, this analysis

indicates a deviation from the 'standard' genetic code, with ATA encoding the start codon methionine, M, instead of isoleucine, I.

All *P.rubra* genes are found on the 'plus' strand. In *I. pulchra*, genes are distributed over both the plus and minus strands, with just two blocks of genes with the same transcriptional polarity clustered together (*rrnL-rrnS-trnI-trnP-trnY*; *trnS2-nad5-trnW-nad2*). Similarly, in *A. ylvae* genes are distributed across the two strands, with two clustered 'blocks' of genes and tRNAs (*nad4l-nad6-nad4-trnT-cox2-nad3-trnV-trnY*; *trnM-trnH-trnQ-trnG-trnS1-trnI-trnF-trnA-trnL1-trnE-trnC*) (Figure 3.1). Whilst the *P. rubra* mitochondrial sequence is condensed, with a large degree of overlap between adjacent genes, the opposite is true for *I. pulchra* and *A. ylvae*. Unlike other metazoan mitochondrial genomes, where genes are adjacent or overlapping and one or two larger non-coding regions are commonly found, *I. pulchra* non-coding sequence is found consistently between protein-coding genes and between tRNAs, ranging in length from twelve to 277 base pairs. In addition, three long non-coding regions of 788, 1143 and 717 base pairs are found at the start of the sequence; between *trnL1* and *trnN*; and *trnN* and *rrnL* (Table 2). The A+T content of these three sections is 68.78%, 65.79% and 76.15% respectively. Of these regions, the compositional difference between the 717 base pair non-coding sequence and the rest of the genome is statistically different (χ^2 = 25.629, p<0.0001), with a higher A+T content indicating that it could function as a transcriptional control region. There is also a large portion of intergenic, non-coding sequence in the *A. ylvae* mitochondrial genome. Eight regions of non-coding sequence greater than 100 base pairs are distributed throughout the genome, with 24 additional smaller intergenic regions, ranging in size from three to 70 base pairs (Table 3). Of the larger non-coding sequences, three have an A+T content that is statistically higher than the entire sequence: 309 nucleotides between *cox1* and *nad4l* (χ^2 = 3.944, p<0.1); 126 nucleotides between *nad4l* and *nad6* (χ^2 = 6.964, p<0.01) and 260 nucleotides between *trnS2* and *rrnL* (χ^2 = 9.654, p<0.01). The *P. rubra* sequence has just one longer non-coding sequence, of 196 base pairs.

The gene arrangement in all three acoel mitochondrial genomes I analysed is unique amongst published metazoan mitochondrial genomes (Figure 3.6). The species analysed in this study share only the small 'block' of *nad3-atp6-nad4-cob* (*I. pulchra*) and *cob-nad4-nad3* (*P. rubra*). However, the order is reversed between the two, and the genes are distributed across both strands in *I. pulchra*, and so it is perhaps unlikely that this represents a feature inherited from a common ancestor. To quantify the number of common gene arrangements between the species in this study and other mitochondrial genomes, I analysed protein-coding gene and ribosomal RNA gene order using CREx[112] (compared to the acoel *S. roscoffensis*, *Xenoturbella bocki* and the metazoan mitochondrial 'ground plan', represented by *L. polyphemus*). tRNAs were not included in analysis owing to their more frequent translocation in mitochondrial genomes. Conserved gene 'blocks', defined as a common series of genes, regardless of their order within that grouping, were very infrequent between the species. Of the genomes compared, the highest number of common gene blocks was found between *X. bocki* and *P. rubra*, but this result was not significant, finding only 16 common intervals out of a possible 150, and confirming the visual observation that gene order between these species is highly variable.

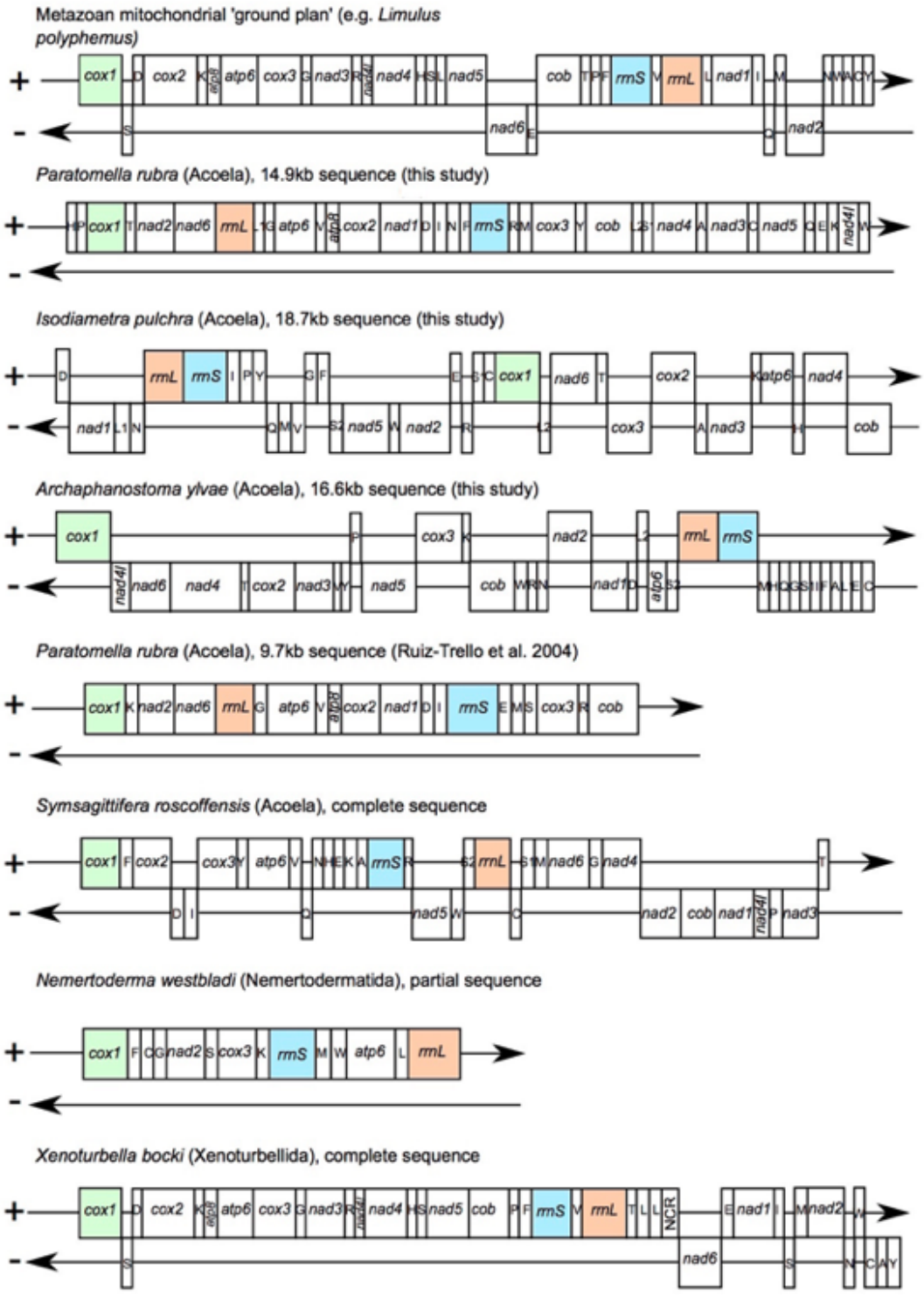


Figure 3.6. Comparison of gene orders in Acoela mitochondrial genome sequences. *P. rubra*, *I. pulchra* and *A. ylvae* genomes from my analysis compared to a published *P. rubra* fragment; the acoel *S. roscoffensis*; the xenoturbellid *X. bocki*; the nemertodermatid *Nemertoderma westbladi* and the metazoan mitochondrial 'ground plan' gene order, represented by *Limulus polyphemus*. Genes are not drawn to scale. Coloured genes chosen

to show 'anchors' and divergence from the ground plan order in other species.

3.2.3 Phylogenetic analysis and population differentiation

The new mitochondrial data from *P. rubra*, *I. pulchra* and *A. ylvae* was used to investigate the internal phylogeny of the acoels and to test support for an Acoela-Xenoturbellida Xenacoelomorpha affinity. First analyses showed that including the fast-evolving tunicates in phylogenetic inference led to a clustering of the tunicates and the acoels in an artificial long-branched clade (Figure 3.7). When the tunicates were removed from analysis, both Bayesian phylogenetic inference and maximum likelihood approaches generated the same topology (Figure 3.8). The protostome/deuterostome split was correctly inferred and Xenacoelomorpha were found splitting off inside the Deuterostomia. *P. rubra*, *I. pulchra* and *A. ylvae* were all grouped inside Acoela, as would be expected. The Nemertodermatida species *Nemertoderma westbladi*, used to represent the nemertodermatids within the Acoelomorpha (=Acoela + Nemertodermatida) branched inside Mollusca, indicating that the limited mitochondrial data available for this species on GenBank is a contamination.

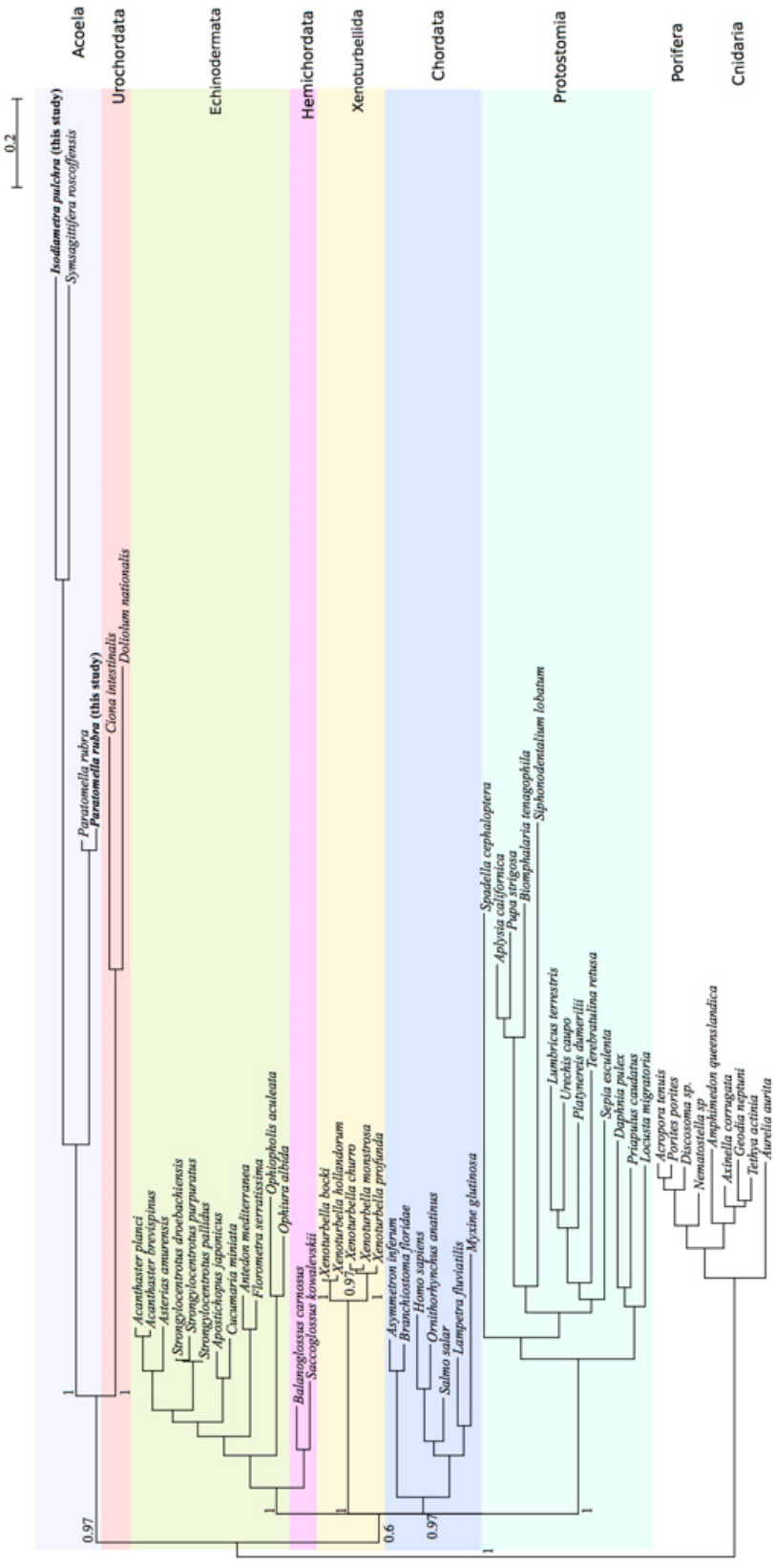
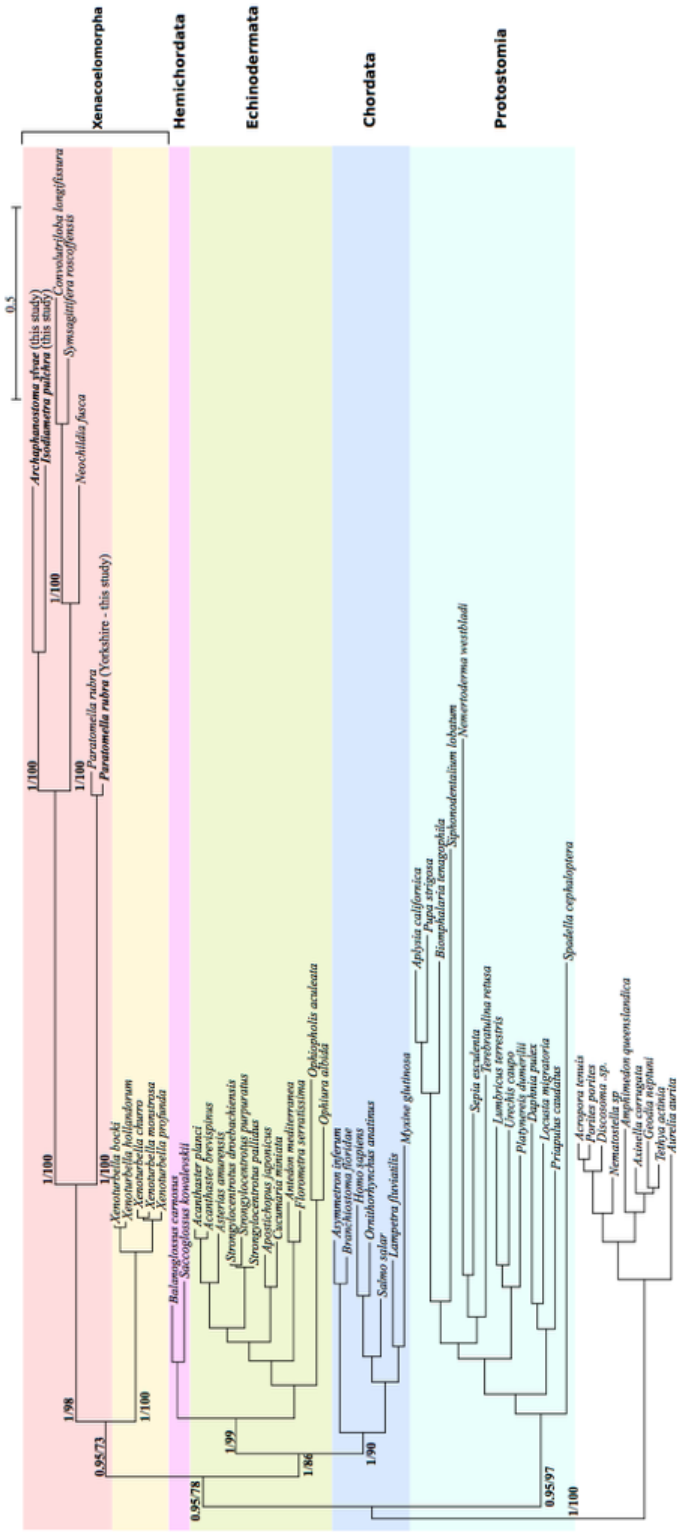


Figure 3.7. Initial Bayesian phylogenetic analysis of mitochondrial protein-coding genes from the Metazoa. Analysis includes initial data from *P. rubra* and *I. pulchra*, with posterior probabilities at relevant nodes. Analysis carried out by Philipp Schiffer (Telford lab) on a trimmed alignment using PhyloBayes. Branch length value shows number of substitutions per site.


Xenacoelomorpha
Hemichordata
Echinodermata
Chordata
Protostomia
0.5
Archaphanostoma ivtae (this study)
Isodiametra pulchra (this study)
Convolutriloba longifissura
Symsagittifera roscoffensis
Neochildia fusca
Paratomella rubra
Paratomella rubra (Yorkshire - this study)
Xenoturbella bocki
Xenoturbella hollandorum
Xenoturbella japonica
Xenoturbella monstrosa
Xenoturbella profunda
Balanoglossus carnosus
Saccoglossus kowalevskii
Acanthaster planci
Asterias amurensis
Strongylocentrotus droebachiensis
Strongylocentrotus purpuratus
Apostichopus japonicus
Cucumaria echinata
Asterina pectinifera
Anneissia japonica
Ophiopholis aculeata
Ophiura albida
Florometra serratissima
Acanthocystis turfacea
Branchiostoma floridae
Homo sapiens
Ornithorhynchus anatinus
Salmo salar
Lampetra fluviatilis
Myxine glutinosa
Aplysia californica
Pupa strigosa
Biomphalaria tenagophila
Siphonodentalium lobatum
Nemertoderma westbladi
Sepia esculenta
Terebratulina retusa
Lumbricus terrestris
Urechis caupo
Platynereis dumerilii
Daphnia pulex
Lepidodasys antarctica
Priapulus caudatus
Spadella cephaloptera
Acropora tenuis
Porites porites
Discosoma sp.
Nematostella sp.
Amphimedon queenslandica
Axinella corrugata
Geodia neptuni
Tethya actinia
Aurelia aurita
1/100
1/100
1/100

The previously published 9.7kb fragment of *P. rubra* mitochondrial genome was derived from a population sampled near Barcelona (Spain). By comparing this sequence to the sequence generated from the individuals I sampled from Yorkshire (UK), the total sequence divergence could be estimated, and non-synonymous to synonymous substitutions for eight protein-coding genes compared. The overlapping 9.7kb sequence was only 82.63% similar at the nucleotide level. The number of substitutions varied between, for example, 23 in the shortest gene alignment (*atp8*; 177 base pairs), to 161 in *nad2* (972 base pairs), and 116 in the 1401 base pair long *cox1* alignment. Interestingly, non-synonymous substitutions are frequent: 13 in *atp8*, 104 in *nad2*, and 25 in *cox1* (Table 4). Furthermore, similarity of the *cox1* sequences at the nucleotide level is only 91% over 666 base pairs, which is lower than the 95-98% threshold used to distinguish species based on *cox1* barcoding[113].

Table 4: Substitution pattern differences between *P. rubra*. Nine genes found on the published *P. rubra* mitochondrial genome (Barcelona, Spain) compared to samples from Yorkshire, UK. S-Sites and N-Sites denote the number of possible substitution sites in the gene that will result in a synonymous or non-synonymous substitution, respectively. For each codon, non-synonymous sites are calculated as the total fraction of possible non-synonymous substitutions at each codon position (1, 2 or 3); synonymous sites are calculated as all possible substitution sites, minus the number of non-synonymous substitution sites. Of these possible substitution sites, S-substitutions and N-substitutions denote the observed substitutions between the two *P. rubra* genomes.

Sequence	Length (bp)	S-Sites	N-Sites	Substitutions	S-Substitutions	N-Substitutions
atp6	594	90.8602	503.14	70	40.9815	29.0185
atp8	177	15.4355	161.565	23	10	13
cob	795	113.09	681.91	89	63.9542	25.0458
cox1	1401	158.705	1242.3	116	90.904	25.096
cox2	660	116.441	543.559	81	50.9427	30.0573
cox3	780	99.2387	680.761	87	60.5296	26.4704
nad1	930	95.8574	834.143	103	63.3614	39.6386
nad2	972	156.176	815.824	161	56.7858	104.214
nad6	330	44.3145	285.685	52	31.121	20.879

3.3 Discussion

3.3.1 Difficulties in resolving the complete circular genome of *P. rubra* and *I. pulchra*

3.3.1.1 Problems in 'closing the circle' of the P. rubra mitochondrial genome

The 14.9kb sequence that I was successfully able to amplify and verify for *P. rubra* contains the full complement of 37 genes that are typical of metazoan mitochondrial DNA. A number of lab-based and computational efforts to close the circle of the mitochondrial genome were attempted, but none of these proved successful. From an experimental set-up perspective, I designed many different 'closing genome' primers within confirmed protein-coding gene sequence to ensure sequence validity. These were tried in numerous forward and reverse primer combinations, with different polymerases; PCR cycling protocols, annealing temperatures; temperature 'touchdown' strategies; and 'nested' PCRs, where previous PCR products were used as a template in subsequent reactions with primers designed within the sequence covered by the previous primers. Computationally, I mapped transcriptome reads at either end of the 14.9kb sequence in the hope of finding the same read mapping at both locations, but no sequencing data was able to resolve the missing region. It is possible that the difficulty found in trying to resolve the circular mitochondrial sequence could be attributed to the very AT-rich, repetitive sequence found at both ends of the fragment, which could have prevented successful PCR amplification. Similar regions have been shown as problematic in studies of other mitochondrial genomes[114]. As no stretch of long non-coding sequence was found for this species in our study, the missing sequence might represent its mitochondrial transcription control region, characterised by a high A+T content. However, the overall AT content of the *P. rubra* mitochondrial sequence (78.15%) is high even for mtDNA, and greater than the A+T content of the mitochondrial genome of the acoel *S. roscoffensis* (75.3%)[19] and the published partial *P. rubra* genome (76.4%)[20].

3.3.1.2 Investigating possible sequence duplication in I. pulchra

The validity of the duplicated sequence found in the *I. pulchra* mitochondrial genome could not be confirmed by PCR or by any computational efforts to map short reads to resolve it. PCR set up using primers flanking the duplicated region were unsuccessful despite using polymerases and following protocols optimised for long-range amplification (>5kb). As no PCR amplicon was able to span the entire duplicated region, including sequence on both sides of the duplicate, it was more reliable to exclude this duplicated sequence from analysis (Figure 3.3). Nonetheless, duplications within mitochondrial genomes are not uncommon, and changes to mitochondrial gene order are widely thought to arise as the result of a sequence 'duplication and deletion' mechanism[99,115,116]. A number of mitochondrial genomes with duplicated sequences have been reported in species with a divergent mitochondrial gene order[116-118]. Given the highly unusual gene order of the *I. pulchra* mitochondrial genome, a genomic duplication could provide evidence for a genomic 'duplication and deletion' rearrangement of genes. The rearrangement and separation of protein-coding genes in other mitochondrial genomes has been attributed to long, non-tandem, inverted repeats[117] – and this could also be true for the *I. pulchra* mitochondrial genome. Furthermore, very long nematode mitochondrial genomes with variable duplicated regions have been described with a conserved region containing the majority of the protein-coding genes[119]. In *I. pulchra*, the protein-coding genes and tRNAs – with the exception of *nad1*, *trnD* and *trnL1* – are found grouped together in one main block, outside of the duplicated section. Despite these factors providing a degree of evidence for the validity of the duplicated sequence, long, non-coding duplications are most commonly found adjacent to tRNAs or alongside other sequences capable of forming stem-and-loop structures. This is not the case for the potential duplicate in *I. pulchra*. Furthermore, both occurrences of the duplicate are identical, nucleotide-by-nucleotide, and unless the duplication occurred exceptionally recently, it is more than likely that spontaneous mutations would result in nucleotide differences between

the two copies of the sequence – especially given the elevated mutation rate of mitochondrial genomes. It is also true that the locations of the duplicated sequences are at the start and end point of transcriptome assembly contigs, meaning that the duplicates could have arisen solely as a result of a sequencing and assembly error. Their existence is nonetheless supported by PCR products, which show an identical sequence being adjacent to both *rrnL* and *cob*. Validation of the presence of a duplicated sequence would be provided by PCR amplification spanning the length of and either sides of both duplicated sequences, but this was not possible, despite numerous PCR strategies.

3.3.2 Divergent gene orders in Acoela mitochondrial genomes

The 14.9kb mitochondrial genome of *P. rubra*, the 18.7kb sequence from *I. pulchra*, and the complete 1.6kb *A. ylvae* mitochondrial genome that I was successfully able to amplify and/or verify show no significant organisational similarity to any other published metazoan mitochondrial genome.

The *P. rubra* sequence I report has an identical protein-coding and ribosomal gene order with the previously published 9.7kb *P. rubra* sequence, but has variation in tRNA order. As tRNAs are reported to show much more frequent gene translocation compared to larger genes[120], this could account for this discrepancy. Nonetheless, the variation in tRNA location, along with the surprising difference at the nucleotide level between the two sequences, could indicate that *P. rubra* collected from Barcelona (Spain) and the animals collected for this analysis (Yorkshire, UK) should be regarded as cryptic species – and not just divergent populations. Given the largely unresolved diversity of benthic communities[121], and the wider marine environment in general[122], differentiation of *P. rubra* in different populations into cryptic species is perhaps not surprising. This finding also highlights the usefulness of studying mitochondrial genomes to understand hidden species diversity.

All three acoel species are unique in both the orientation and orders of their genes: *P. rubra* has genes transcribed exclusively in one orientation, on one strand whilst *I. pulchra* has an almost-equal distribution of genes across both strands (18 genes vs.17 genes). With *cox1* in a forward orientation at the start of the genome (as is convention for mitochondrial genomes), the majority of the protein-coding genes for *A. ylvae* are found on the minus strand (Figure 3.6). Genes in both *I. pulchra* and *A. ylvae* are not grouped together into long gene blocks of the same transcriptional orientation, but are found distributed as one or two genes on each strand. The unique order and distribution of genes for these species seems to be typical for the acoels: analysis of the complete *S. roscoffensis* mitochondrial genome found no gene order similarity to any other species published to date[19]. It is possible that the variability in mitochondrial gene order within the Acoela - and compared to other taxa - could be a consequence of the rapid rate of sequence change observed for this lineage. Although data for this group is limited to just four species, the uniqueness of acoel mitochondrial genomes analysed so far, and the absence of any characteristic 'gene blocks' means that gene order may not be phylogenetically informative for the Acoela. Further mitochondrial genome data from other members of the Acoela would no doubt aid in this comparative analysis.

3.3.3 Gene overlap and non-coding DNA

The mitochondrial genome of *P. rubra* shows frequent overlaps between protein-coding genes and tRNAs. tRNAs have been reported within protein-coding genes in other metazoan mitochondrial genomes[123,124], and given that no other location could be predicted for these sequences, the observed overlap could represent the simultaneous coding for both tRNAs and protein-coding genes. Overlap in coding sequence could be the result of selection to reduce genome size, accompanied by a reduction in non-coding sequence[124], and truncated tRNAs with incomplete secondary structure – both of which are also found for the *P. rubra* mitochondrial sequence.

Interestingly, the opposite is true for the *I. pulchra* and *A. ylvae* sequences. For *I. pulchra*, the sequence that I could confidently verify makes the minimal possible length of the *I. pulchra* mitochondrial genome 18,725 nucleotides, and it is likely to be longer in the complete closed circular genome, whether the duplicated sequence is valid or not. As in other 'long' mitochondrial genomes, this increased length is largely due to an increase in non-coding stretches of DNA[125]. Indeed, the lengths of protein-coding genes inferred for *I. pulchra* are similar to those of other acoel species (Table 5), and two protein-coding genes have been lost from the genome, contributing to a reduced proportion of protein-coding gene sequence within the genome. The loss of *atp8* is not unusual, and this has been reported in a number of unrelated taxa, as well as in *S. roscoffensis* and *A. ylvae* within the Acoela[19]. The absence of *nad4l* in *I. pulchra* is more unusual, and it is possible that this gene exists in a portion of the genome that I could not successfully sequence or verify. Although non-coding sequence contributes a relatively large proportion of the *A. ylvae* mitochondrial genome (17.1% compared to 22.72% in the *I. pulchra* sequence), the total genome is not exceptionally long.

Table 5: Length of protein-coding genes in acoel mitochondrial genomes. All gene lengths in base pairs.

	Isodiametra pulchra	***Paratomella rubra***	***Symsagittifera roscoffensis***	***Archaphanostoma ylvae***
cox1	1536	1563	1551	1539
cox2	615	663	741	648
cox3	798	786	792	786
nad1	881	1053	870	870
nad2	1053	1014	990	1002
nad3	378	390	393	369
nad4	1344	1326	1350	1344
nad4l	absent	309	270	285
nad5	1710	1752	1776	1656
nad6	476	462	480	468
cob	1134	1083	1161	1086
atp6	681	609	702	702
atp8	absent	177	absent	absent

3.3.4 Phylogenetic inference using mitochondrial data

The internal phylogeny resolved for Acoela is in line with that proposed by Jondelius *et al.* (2011)[126]. *I. pulchra* and *A. ylvae* group together in the Isodiametridae; Isodiametridae groups with *S. roscoffensis*, *Neochildia fusca* and *Convolutriloba longifissura* (the latter two represented by *cox1* data only), which are all members of the Convolutidae; and *P. rubra* forms a separate branch outside the Convolutidae, representing the Paratomellidae. The initial grouping of the acoels and tunicates is likely to be a classical example of LBA, resulting from the fast sequence evolution of mitochondrial DNA and compounded by the long-branched acoel members (Figure 3.7). The accelerated substitution rates in mitochondrial DNA are also evidenced by the cryptic divergence we find in *P. rubra*, and may well lead to LBA in phylogenies derived from mitochondrial protein-coding genes, owing to the

clustering of rapidly evolving lineages. Again, this is of particular relevance for acoel species, which already demonstrate a very rapid rate of nucleotide substitution compared to other metazoans, leaving them even more vulnerable to LBA. When the Urochordata are excluded from analysis, Xenacoelomorpha are resolved as a branch within the deuterostomes (Figure 3.8), as has been found in other phylogenies derived from mitochondrial gene sequences[23,127]. Nonetheless, this analysis clearly emphasises the problematic placement of Xenacoelomorpha in molecular phylogenies. In particular, the long-branched Acoela are clearly drawn to the base of the tree in LBA when the urochordates are included in analysis, highlighting the problem of systematic error in phylogenies derived from these taxa.

3.4 General conclusions

Starting from genome assemble contigs, I was successful in sequencing and verifying one complete mitochondrial genome (*A. ylvae*), and two partial mitochondrial genomes (*P. rubra* and *I. pulchra*) from the Acoela. Despite being unable to close the circle of the mitochondrial genome for *P. rubra* and *I. pulchra*, the data from these species comprises all protein-coding genes for *P. rubra*, and all eleven of the protein-coding genes that are likely to be found in the *I. pulchra* mitochondrial genome. The addition of mitochondrial genomes from these three species increases the relatively sparse molecular data available for the Acoela: prior to this work, only one complete mitochondrial genome had been published.

It is known that the Acoela are long-branch, rapidly evolving taxa, and the divergent gene orders found in the mitochondrial genomes of these taxa could be a result of this. It is also evident that phylogenetic inference using molecular sequence data from the Acoela is likely to be hindered by LBA if not properly accounted for. Increased sampling from across the Acoela to identify more slowly evolving representatives (along with *P. rubra*) could go some way to help 'break' the long branch. Similarly, more data from

mitochondrial genomes of early branching taxa could help to better-inform phylogenetic inference.

4 Molecular markers of excretion: conservation of ultrafiltratory genes

4.1 Introduction

4.1.1 The Nephrozoa and Xenacoelomorpha

As outlined in section 1.3, the simple organisation of the Xenacoelomorpha means that they are commonly assumed to lack any cells or systems specialised for an ultrafiltratory or excretory function (nephridia). This assumed absence of nephridia has been used as evidence for their exclusion from the main protostome and deuterostome grouping, which has, in consequence, been termed the Nephrozoa.

Despite this widely held assumption, to date no structural or functional assays have been used to date to investigate excretion in any member of the Xenacoelomorpha. Given the disparity in form of nephridial systems found across the Bilateria, identifying common morphological or molecular characteristics of nephridia would be a useful approach for investigating the presence of such specialised structures in *Xenoturbella* or the Acoela.

To better understand the homology of different nephridial systems, we can examine conservation between taxa within discrete regions of the overall system. In recent years, an increasing amount of investigation has focused on the site of ultrafiltration in different model taxa, namely the vertebrates, *D. melanogaster*, and *S. mediterranea*. In next two sections I outline the evidence for morphological and molecular conservation at the site of ultrafiltration in representatives from the three main bilaterian groupings: the Deuterostomia, Lophotrochozoa, and Ecdysozoa.

4.1.2 Morphological conservation at the site of ultrafiltration

Investigation into the site of ultrafiltration in the vertebrate kidney and the *D. melanogaster* nephrocyte in recent years has provided evidence for conserved morphology: an initially surprising finding given that both structures appear at first glance to be unique examples of very modified metanephridia.

4.1.2.1 The vertebrate podocyte

In the glomerulus of the vertebrate kidney specialised foot-shaped epithelial cells (podocytes) derived from the intermediate mesoderm carry out ultrafiltration along with the glomerular basement membrane (Figure 4.1). The 'toes' (pedicels) of podocytes surround the adjacent glomerular capillary. Foot processes are separated by slits of 30-50nm in diameter, spanned by a so-called 'slit diaphragm' which lines up with the fenestrations in the underlying capillary endothelial cells[128]. The podocyte slit diaphragm and the glomerular basement membrane (GBM), an extracellular matrix component which lies between the endothelial cells and the podocyte epithelial cells, are the only barriers between blood in the glomerular capillary and the inside of the Bowman's capsule. These two structures together act as a molecular filter: proteins larger than ~10kDa are excluded from passing into the Bowman's capsule, but the passage of water, ionic compounds, sugars, amino acids, hormones and various nitrogenous waste products is permitted. Both the GBM and the podocyte foot processes are also negatively charged, preventing the passage of other negatively-charged molecules, including most plasma proteins[129].

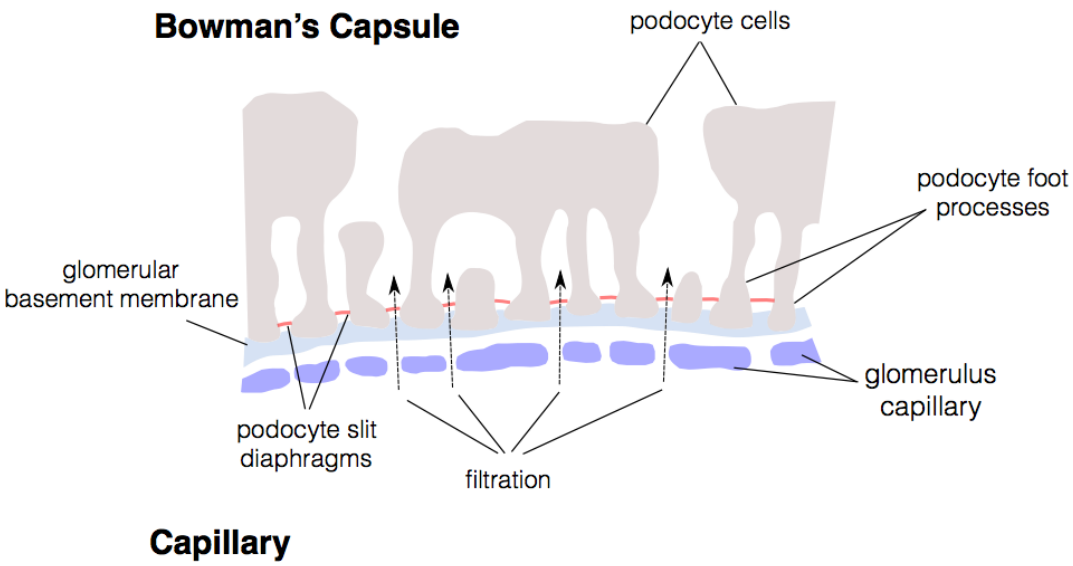


Figure 4.1. Ultrafiltration mediated by podocytes into the Bowman's capsule in the vertebrate glomerulus. Filtration occurs across three 'barriers'. (1) Fenestrations of between 50-100nm in the capillary of the glomerulus (shown in purple) allow the passage of fluid, proteins and other plasma solutes. (2) Primary filtrate crosses the thick glomerular basement membrane (GBM, shown in blue). The 250-400nm thick negatively charged GBM functions to filter solutes on their size and charge. (3) Specialised epithelial cells called podocytes (grey) have cytoplasmic foot processes that enwrap the adjacent glomerular capillary. The space between foot processes is spanned by a slit diaphragm (red) – formed by a network of proteins – which functions as the final stage of ultrafiltration into the Bowman's capsule. Slit diaphragms are 30-50nm in diameter and prevent the passage of solutes larger than ~10kDa into the Bowman's capsule. Foot processes are also negatively charged, enhancing the charge-based filtration occurring across the GBM.

4.1.2.2 *Drosophila* nephrocytes

In *D. melanogaster*, ultrafiltration is facilitated by specialised cells called nephrocytes, which sequester and/or metabolise waste compounds from the haemolymph[130] (Figure 4.2). Two types of nephrocytes are found in the larval *D. melanogaster*: pericardial nephrocytes, found in rows of 20-25 cells on either side of the heart, and garland-cell nephrocytes, present as a fused 'necklace' around the oesophagus (Figure 4.2A)[131,132]. Most larval nephrocytes persist through metamorphosis into the adult *D. melanogaster*, where they are characterised as either thoracic or abdominal.

As well as having the same ultrafiltratory function, there are a number of other striking similarities between *D. melanogaster* nephrocytes and vertebrate podocytes (Figure 4.2B) [130].

1) Similar to the vertebrate podocyte, the plasma membrane of the fly nephrocyte is extensively infolded to form a series of lacunae, flanked on either side by nephrocyte foot-processes.
2) In flies, as in vertebrates, the entrance to these channels is a slit of ~30nm width, spanned in it's entirety by a single or double filament (the nephrocyte diaphragm).
3) The fly nephrocyte is enveloped by the negatively charged basement membrane, and this, along with the nephrocyte diaphragm, functions as a size- and charge-based filtration barrier to molecules in the haemolymph, preventing the passage of molecules larger than ~10kDa[133].

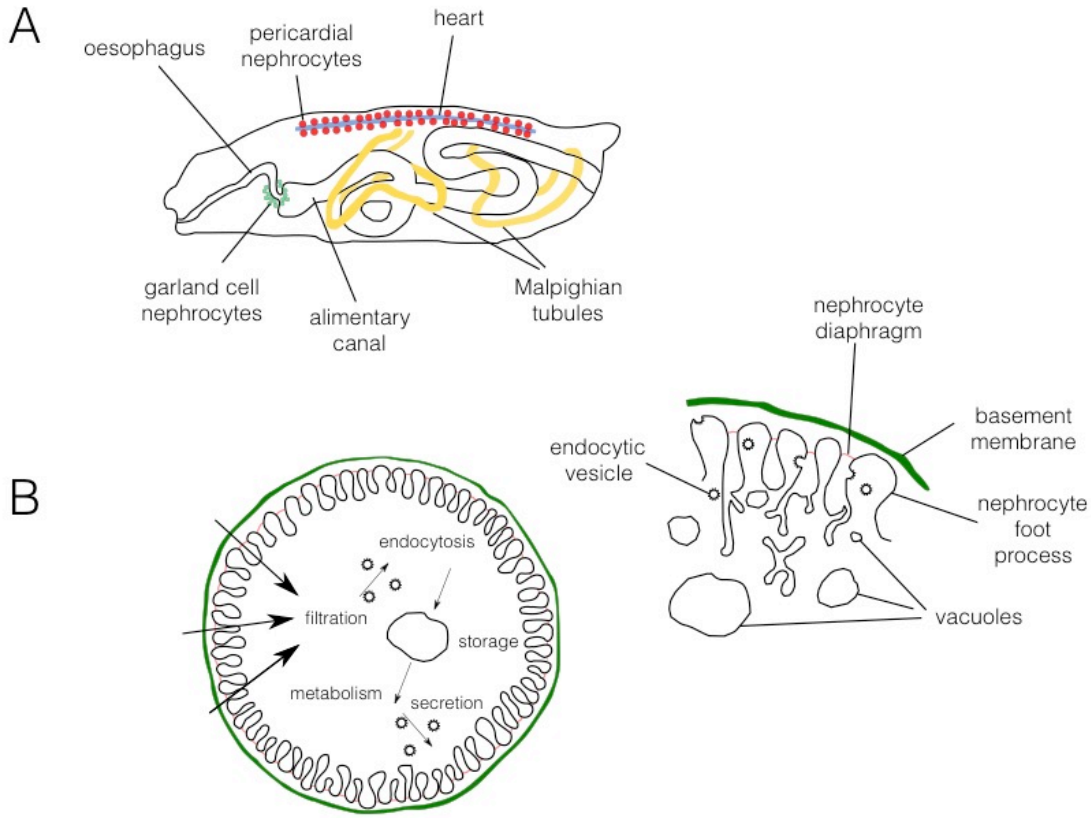


Figure 4.2. Ultrafiltration in the nephrocytes of *D. melanogaster*. Figure adapted from Weavers *et al.* (2009)[130] (A) Distribution of nephrocyte cells – the site of ultrafiltration in *D. melanogaster*. Pericardial nephrocytes (red) found in a row either side of the heart (blue). Garland cell nephrocytes (green) found as a 'necklace' around the oesophagus. Malpighian tubules, shown in yellow, are the site of osmoregulation and filtrate modification. They are found as two pairs of tubules, connected to the gut via ureters. (B) *D. melanogaster* nephrocyte. Left: haemolymph is filtered across the basement membrane (outer green border) and the nephrocyte diaphragm (red line between infoldings) and endocytosed. Filtered material is stored in vacuoles and/or metabolised and secreted back into the haemolymph. Right: detail of filtratory apparatus in the nephrocyte.

4.1.2.3 Ultrafiltration in the flame cells of protonephridia

In the platyhelminth *S. mediterranea*, ultrafiltration occurs in a scattered population of so-called 'flame cells' distributed across the body. Similar to the podocyte and nephrocyte, flame cells have foot processes 90-150nm wide, between which there are slit-like fenestrations 35-40nm wide which, along with the ECM, function as the ultrafiltratory barrier[37]. Unlike vertebrates, planarians lack a circulatory system, and cannot rely on pressure-driven flow for fluid ultrafiltration. Instead, ultrafiltration in *S. mediterranea* is facilitated by a ciliary-mediated current: vigorous beating of cilia in the flame cells (resembling the flickering of a flame) draws body fluid across the terminal region of the protonephridia, where ultrafiltration occurs[43] (Figure 1.10).

4.1.3 Molecular conservation at the site of ultrafiltration

In addition to the apparent morphological conservation, there also appears to be molecular conservation of the structural and signalling proteins necessary to form the filtratory apparatus in diverse bilaterian taxa (Figure 4.3). Given the diverse morphology of nephridial systems, the identification of shared molecular expression at the site of ultrafiltration is an important finding for understanding the evolutionary origin and homology of nephridial systems.

In the vertebrate podocyte, two structural proteins have been shown to be necessary for the formation of the slit diaphragm: the cell-adhesion molecules (CAMs) Nephrin and Neph1, belonging to the Immunoglobulin superfamily of proteins (IgSF)[134]. Mutations to the genes coding for Nephrin and Neph1 proteins, *NPHS1*(/*Nephrin*) and *NEPH1*(/*Neph1*) respectively, are causal for a number of diseases relating to abnormal renal function in humans. Nephrin and Neph1 are co-expressed only in the podocytes, where their extracellular domains interact to make heterotypic (Nephrin-Neph1) and homotypic (Nephrin-Nephrin) dimers and multimers[135]. These structures form

the physical and molecular basis of the slit diaphragm itself, and determine the permeability of the slit diaphragm to molecules passing into the Bowman's capsule[134-137].

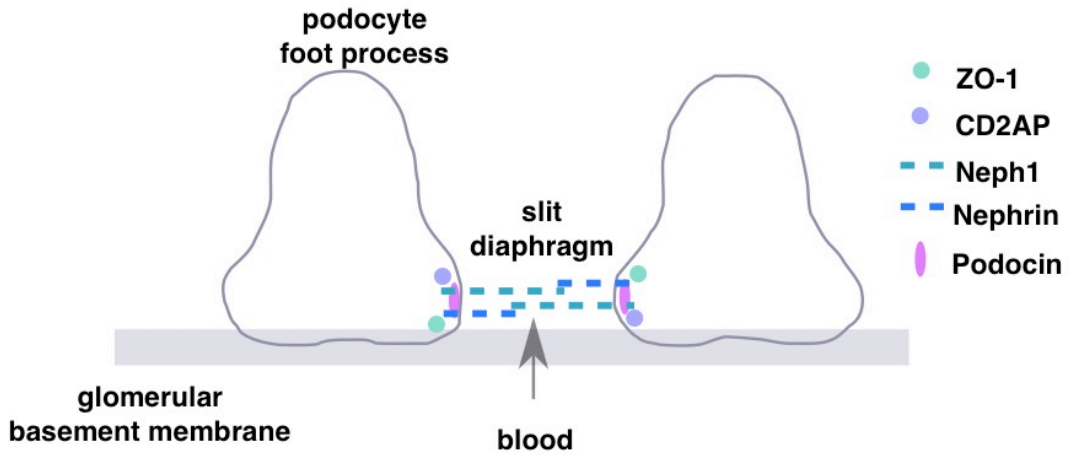


Figure 4.3. Formation of the podocyte slit diaphragm via the interaction of structural proteins. The formation of the slit diaphragm itself is dependent on homodimers and heterodimers formed between the structural proteins Nephrin and Neph1 (blue dashed and green dashed lines respectively). Neph1 binds ZO-1 (green circle) for protein organisation and signal transduction. CD2AP (purple circle) binds Nephrin and is necessary for structural maintenance of the slit diaphragm: CD2AP knockout mice die of renal failure at six or seven weeks. Podocin (pink oval) interacts with CD2AP and Nephrin for structural and functional support of the slit diaphragm. This is not an exhaustive overview of all structural proteins and signalling molecules involved in slit diaphragm formation; instead, those with a critical function and for which orthologues have been identified across the Bilateria are shown.

Two orthologues each of the genes *NPHS1* and *NEPH1* have been identified in *D. melanogaster*: *sticks and stones* (*Sns*) and *hibris* (*Hbs*) for *NPHS1*, and *dumbfounded* (*Duf* – also know as *Kirre*) and *roughest* (*Rst*) for *NEPH1*[130]. Of these, the proteins Sns and Duf are found in both garland and pericardial nephrocytes, and the onset of their expression (from mid-embryogenesis in garland nephrocytes and from the first larval instar in pericardial nephrocytes) correlates with the timing of the first appearance of the nephrocyte diaphragm[138]. The only cell type in which Sns and Duf are co-expressed is the nephrocyte, where they co-localise specifically to the nephrocyte diaphragm and form heterodimers in trans, mirroring the co-expression of Nephrin and Neph1 in the vertebrate podocyte[130,139]. Sns and

Duf are mutually dependent on each other for stabilisation at the plasma membrane: knock-out of either protein results in the loss, reduced expression, or mis-expression of the other. In flies with mutated forms of *Sns* and/or *Duf*, the nephrocyte diaphragm fails to form, and the number of lacunal infoldings is reduced: where lacunae do form, the nephrocyte diaphragm is always missing[130]. Similarly, in knock-down experiments for Sns (*Sns*) and Duf (*Duf*), the number of nephrocyte diaphragms that form is dramatically reduced, and size-based filtration of molecules is impaired[139]. These phenotypes are the same as those observed in the mutation or absence of Nephrin and Neph1 in the vertebrate podocyte, indicating that the critical molecular components of ultrafiltration and the podocyte slit diaphragm/nephrocyte diaphragm are conserved between vertebrates and insects.

The interaction of Nephrin and Neph1 at the vertebrate slit diaphragm acts as a scaffold for a multi-protein complex involving at least three other proteins. Vertebrate Neph1 binds zonula occludens (ZO-1), a PSD95/Dig/ZO01 (PDZ) domain-containing protein, which is thought to organise Neph1 proteins and facilitate Neph1 signalling via the recruitment of signal transduction components to the slit diaphragm[140]. Similarly, CD2AP localises to the slit diaphragm of the foot processes and binds to Nephrin[141,142]. This is primarily to anchor Nephrin to the cytoskeleton, but the presence of CD2AP also appears to be necessary to maintain the structural integrity of the slit diaphragm: mice lacking CD2AP suffer proteinuria from two weeks of age and eventually die of renal failure by six or seven weeks[142]. Lastly, the hairpin-like protein Podocin is found exclusively in the slit diaphragm of the podocytes where it interacts with CD2AP and Nephrin to maintain the structure and filtratory function of the slit diaphragm, and to facilitate Nephrin signalling[142-144]. *D. melanogaster* has orthologues of all three molecular components (*Pyd* (*ZO-1*), *CG31012* (*CD2AP*) and *Mec2* (*Podocin*), and *in situ* hybridization shows that all three of the genes for these proteins are expressed in fly nephrocytes[130]. More specifically, the intracellular domain of Sns (=Nephrin) has been shown to interact with Mec-2 (=Podocin), and Duf (=NEPH1) to interact with Pyd (=ZO-1). These

interactions are the same as those occurring in the vertebrate podocyte slit diaphragm, providing further compelling evidence for the molecular and structural homology of the nephrocyte diaphragm and the podocyte slit diaphragm.

Recent analysis has also identified 7 *Nephrin* homologues and 3 *Neph1* homologues in the *S. mediterranea genome*: of these, one protein product of each - Smed-NPHS1-6 and Smed-NEPH-3 - localise to the flame cell, although interaction between the protein products of these genes is yet to be determined[37]. Nonetheless, RNAi targeting either gene results in the total absence of the filtration diaphragm, effacement of the lacunae, and the loss of filtratory capability as confirmed by dextran injection and visible bloating of the animal[37]. The apparent conservation between vertebrates, insects and flatworms of molecular components necessary for ultrafiltration is compelling evidence for the common origin of ultrafiltratory cells despite the very different structures and arrangements of their ultrafiltratory apparatuses on a larger organisational scale.

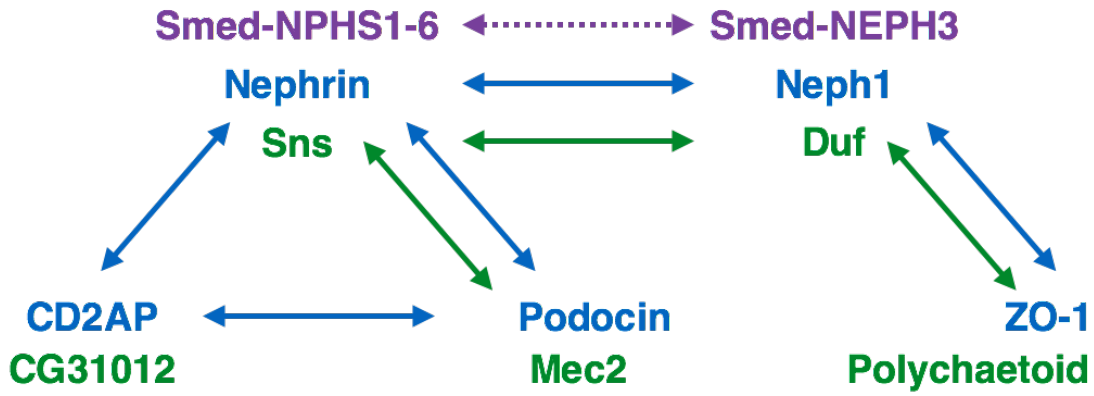


Figure 4.4. Interaction of orthologous proteins at the site of ultrafiltration. Vertebrate proteins shown in blue; *D. melanogaster* orthologous proteins shown in green and *S. mediterranea* orthologous proteins shown in purple. Dashed arrows between *S. mediterranea* orthologues represents the lack of experimental evidence for homodimer and/or heterodimer formation between the Smed-NPHS1-6 and Smed-NEPH3 proteins in the flame cell.

4.1.4 Objectives of chapter

It is clear that all nephridial systems across the Bilateria can be defined by their function of ultrafiltration and osmoregulation. In protonephridia and metanephridia, we consistently find the same two homologous units: that is, a site of ultrafiltration and a tubule element required for filtrate modification and osmoregulation. Furthermore, it is now clear that homologous proteins (Neph1 and Nephrin) are required to facilitate ultrafiltration across the Bilateria. We can also find evidence for similarities in the molecular components of tubule reabsorption and osmoregulation (that is, aquaporins and solute carriers), but these will not be discussed in further detail[36-38].

Despite the main bilaterian grouping (of Protostomia and Deuterostomia) being described as the 'Nephrozoa', some taxa within this grouping lack what would traditionally be classified as nephridia (see 1.3.6). Further to this, if our definition of a 'nephridial system' requires both ultrafiltration and osmoregulation to be carried out, then the simple excretory systems present in both the tunicates and nematodes do not meet this

criterion. Consequently, the current grouping of Nephrozoa is perhaps misleading in its aim of encompassing all animals with filtratory and osmoregulatory capacity.

As has been discussed, there are some bilaterian molecular markers of ultrafiltration and osmoregulation. Whilst osmoregulatory genes – namely aquaporins and solute carriers – are expressed in numerous organ systems and tubule elements throughout the Bilateria, the ultrafiltratory components of Neph1 and Nephrin proteins appear to be co-expressed exclusively at the site of ultrafiltration in vertebrates, *D. melanogaster* and *S. mediterranea*. We know that these genes function more widely in other roles, but whether they had an ancestral role in ultrafiltration, or if this is a 'co-opted' function in nephrozoan taxa remains to be confirmed.

In this chapter, identification of orthologues of these genes in members of the Xenacoelomorpha was undertaken as an initial investigation into the identification of putative markers of ultrafiltration. More widely, I also searched for these genes in diploblasts and non-metazoans, and in bilaterians that are thought to lack any ultrafiltratory capacity, to better understand their distribution across the Eukaryota.

4.2 Results and Discussion

Transcriptomes from whole organism RNA-Seq data from *Xenoturbella*, *P. rubra* and *I. pulchra* were assembled using Trinity[145]; transcriptomic sequences from *S. roscoffensis* were provided by Pedro Martinez. BLAST queries were used to search these transcriptomes for putative orthologues of five ultrafiltratory-related genes (*Neph1*, *Nephrin*, *Podocin*, *ZO-1* and *CD2AP*). Publicly available data for taxa from across the Eukaryota were also searched for orthologues of the same genes. For all genes-of-interest, orthologous sequences were aligned for maximum likelihood phylogenetic inference. Where no orthologous sequence was found for a given species, the best-hit sequences identified using BLAST

were included for analysis. Protein sequences from related protein families were also included as outgroups.

The results of this transcriptome and genome mining are discussed as follows, with each protein or protein class addressed separately. The identification of orthologues, and the known function of each protein in the taxa in which they are found, are discussed therein.

4.2.1 Neph1/Nephrin orthologues and gene function

4.2.1.1 Neph1 and Nephrin are found only in the Bilateria

Neph1 and Nephrin belong to the immunoglobulin superfamily (IgSF) of cell adhesion molecules (CAM). Both Neph and Nephrin proteins have been well conserved throughout evolution: all members share extracellular immunoglobulin-like domains and a short cytoplasmic tail that contains a number of signalling motifs[146-148]. More widely, IgSF proteins carry out a huge number of developmental processes via the formation of heterophilic and hemophilic interactions with other IgSF molecules. Along with Nephrin and Neph protein members, the IgSF CAM class of molecules also comprises neural cell adhesion molecules (NCAMs), plexins, integrins and neurexins, amongst others[149,150].

Annotated sequences for Neph1(=Duf) and Nephrin(=Sns) were available on NCBI for all bilaterian organisms investigated (see 2.6). No annotated orthologues of Neph1 or Nephrin were identified in any non-bilaterian taxa, including diploblastic organisms and representatives from the Choanoflagellatae, Amoebozoa and Filasterea. The best-hit sequences identified in diploblastic organisms using Neph1 and Nephrin as a query were included for phylogenetic inference. These do not group with either Neph1 or Nephrin (Figure 4.5). Orthologues of both Neph1 and Nephrin were found in all representatives of the Xenacoelomorpha included in analysis (Figure 4.5). As shown in Figure 4.5, best hit sequences from Cnidaria (*Acropora digitifera*, *Nematostella vectensis*, *Hydra vulgaris*), Porifera (*Amphimedon*

queenslandica) and Ctenophora (*Mnemiopsis leidyi*) – which were commonly annotated as Hemicentin - group separately from Neph1 and Nephrin. The inclusion of the Nephrin sequence from *Crassostrea gigas* within this outgroup could suggest a mis-annotation of this protein on NCBI.

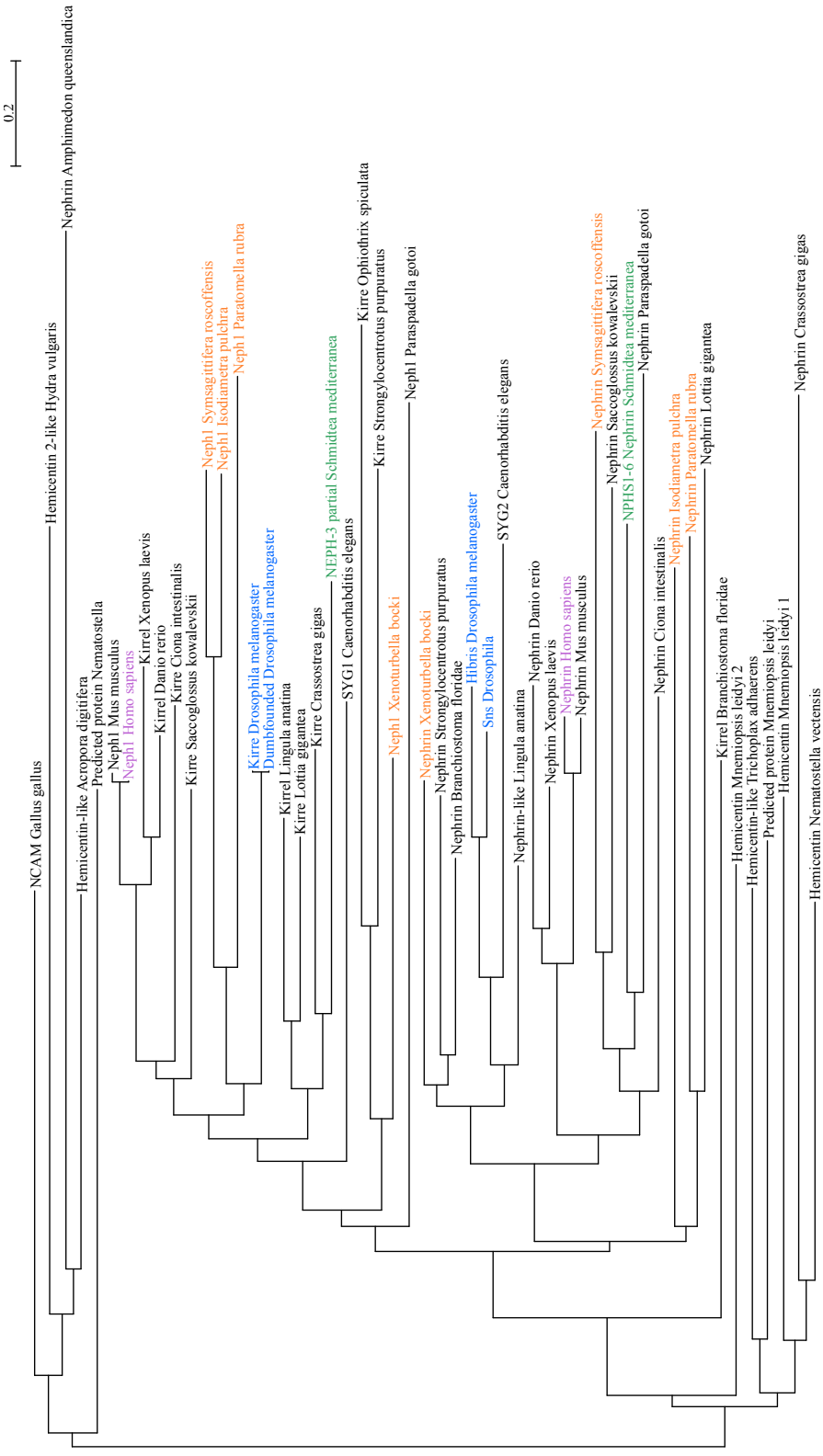


Figure 4.5. Maximum likelihood analysis of the CAM family proteins Neph1 and Nephrin. Xenacoelomorpha sequences shown in orange; *D. melanogaster* sequences in blue (*Kirre*/*Duf* (=Neph1) and *Hibris* and *Sns* (=two orthologues of Nephrin); *H. sapiens* sequences in purple; *S. mediterranea* in green. Phylogenetic inference carried out using RAxML. Branch length value shows number of substitutions per site.

4.2.1.2 Structure and function of Neph1 in the Bilateria

Neph1 is a transmembrane protein characterised by five extracellular immunoglobulin-like domains (Ig-like) (Figure 4.6A). All Xenacoelomorpha orthologues of Neph1 have the same protein domain arrangement as found in vertebrate Neph1 sequences (Figure 4.6). These domains are well conserved in Neph1 protein sequences across the Bilateria. Immunoglobulin-like domains from Neph1 orthologues in *Xenoturbella, P. rubra*, *I. pulchra* and *S. roscoffensis* were aligned to the *H. sapiens* and *M. musculus* protein sequences, and show high similarity to their respective vertebrate Ig-like domains (Figure 4.6B). In Ig-like domains 2-5, all sequences have the common feature of two conserved cysteine residues, separated by ~50 amino acids. These cysteine residues form disulfide bonds within the Ig motif and are defining characteristics of Ig-like domains[151]. Of the four Ig-like domains shown in Figure 4.6B, Ig-like 2 appears to be the most divergent in the Xenacoelomorpha. In particular, in *S. roscoffensis* this domain is longer than those found in the vertebrates and other xenacoelomorphs, and has additional amino acids between conserved residues. Ig-like domain 1 was not well aligned for these species.

Neph orthologues are found throughout the Bilateria – as demonstrated in this analysis - and functional investigation of proteins have been carried out in several bilaterian model organisms. Three Neph paralogues (*NEPH1*, *NEPH2* and *NEPH3*) are found in the vertebrates, expressed as a number of different isoforms with a number of different roles. Diversification of the Neph family in vertebrates is thought to have occurred via duplication events just prior to or during the diversification of the early Vertebrata[147]. Both Neph1 and Neph2 proteins have been found expressed in the vertebrate podocyte[147]. The expression of Neph1 is critical for formation of the podocyte slit diaphragm, and deletion of *Neph1* results in severely defective glomeruli and premature death in mice[152]. It appears that Neph2 also functions in podocyte structure and function, but its role has not been fully characterised, and does not appear to be as critical for

ultrafiltratory function[147,153]. *D. melanogaster* has two *NEPH1* orthologues, *Duf* (also known as *Kirre*) and *Roughest* (*Rst*), which function in a number of different roles. Importantly, the expression of Duf is necessary for specification and function of the nephrocyte diaphragm, as is found for the vertebrate counterpart[130]. Similarly, *NEPH1* orthologues in the platyhelminth *S. mediterranea* are expressed in protonephridia at the site of ultrafiltration. Interestingly, in the nematode *C. elegans*, the *NEPH1* orthologue *SYG-1* functions exclusively in synaptic assembly[154]. No Neph orthologues have previously been identified in any non-bilaterian taxa, and in this analysis no Neph-like orthologues were found in any diploblast or non-metazoan taxa. Instead, the best-hit sequences from blast searches using Neph1 and Nephrin as query sequences against non-bilaterian taxa group outside of the Neph and Nephrin nodes.

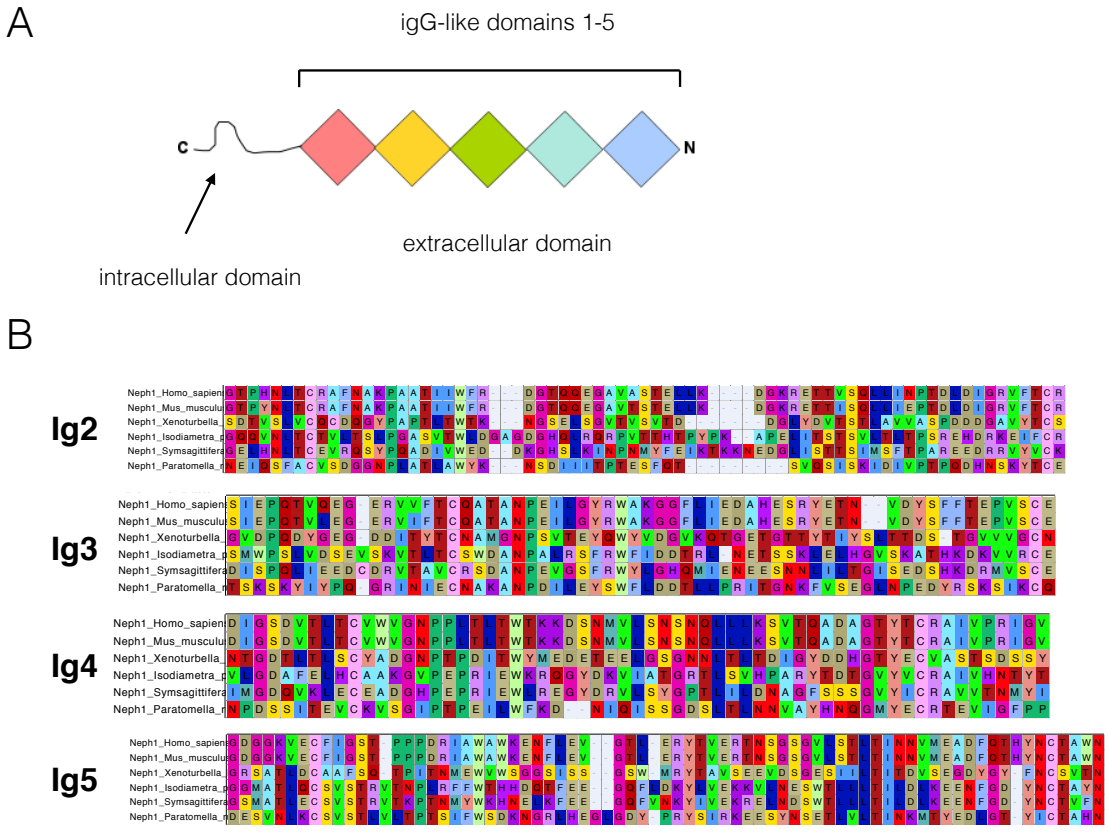


Figure 4.6. Conserved domains in the bilaterian Neph1 sequence. (A) Structure of Neph1, comprising five immunoglobulin-like (Ig) domains in its extracellular region. Schematic based on published vertebrate Neph1 sequences. All Xenacoelomorpha orthologues of Neph1 have the same protein domain arrangement as illustrated here. (B) Conservation of Immunoglobulin-like domains 2-5 between Xenacoelomorpha, *H. sapiens* and *M. musculus*.

4.2.1.3 Structure and function of Nephrin in the Bilateria

Nephrin (encoded by *NPHS1*) is a well characterised transmembrane protein containing eight extracellular immunoglobulin domains, a fibronectin type III-like repeat, and one short intracellular domain[146,155] (Figure 4.7). All Xenacoelomorpha orthologues of Nephrin had the same protein domain arrangement as that illustrated by the vertebrate Nephrin sequence (Figure 4.7).

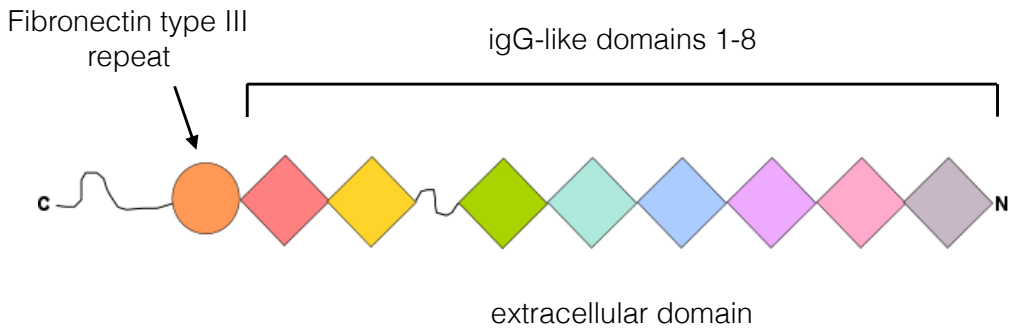


Figure 4.7. Schematic of conserved domains in the bilaterian Nephrin sequence. Domains comprise an intracellular region, a fibronectin III-like repeat and eight conserved Ig-like domains. Schematic based on published vertebrate Nephrin sequences. All Xenacoelomorpha orthologues of Nephrin identified in this analysis have the same protein domain arrangement as illustrated here.

The conserved immunoglobulin and fibronectin-like domains that characterise the bilaterian Nephrin protein sequence are also found in the Xenacoelomorpha sequences (Figure 4.8). As for the Neph1 Ig-like domains, Nephrin Ig domains 2-8 in the Xenacoelomorpha orthologues share two conserved cysteine residues in common with the *H. sapiens* and *M. musculus* sequences. The fibronectin-like domain is well conserved between *H. sapiens*, *M. musculus, P. rubra* and *Xenoturbella*, whilst the sequence from *I. pulchra* appears to contain three unconserved cysteine residues. The longest *Nephrin* transcript identified in *S. roscoffensis* appears to be truncated and did not include the fibronectin-like domain.

Figure 4.8. Alignment of conserved domains in Nephrin orthologues. Immunoglobulin-like domains and a fibronectin III-like repeat in Nephrin orthologues identified in the Xenacoelomorpha. Xenacoelomorphs sequences aligned to human and mouse Nephrin sequences.

No Nephrin-like orthologues have been identified in any non-bilaterian taxa to date, and, as with Neph1 sequences, this was also confirmed in this analysis. Unlike the Neph group of proteins, vertebrate Nephrin is present as just two isoforms, both of which are reportedly expressed in the developing and newborn brain, and one of which localises to the podocyte slit diaphragm[156]. Two *NPHS1* orthologues, *Hibris* (*Hbs*) and *Sticks and Stones* (*Sns*) are found in *D. melanogaster*. Much like Duf and Rst, Hbs and Sns have a number of roles in *D. melanogaster*, but the presence of Sns in the nephrocyte is again necessary for ultrafiltratory structure and function[130]. In *S. mediterranea*, one Nephrin orthologue, Smed-NPHS1-6, localises to the protonephridial flame cell[37]. In *C. elegans*, as shown for SYG-1, the only known expression of SYG-2 is in the synapses[154].

4.2.1.4 The presence of Neph1/Nephrin orthologues in Xenacoelomorpha and inferring ancestral function

As described above, putative Nephrin and Neph1 sequences identified from Xenacoelomorpha transcriptomes group with their respective protein families (Figure 4.5) and have well conserved Ig-like domains (Figure 4.6B and Figure 4.8). The identification of both of these proteins in acoels and *Xenoturbella* supports previous suggestions that Neph and Nephrin proteins are bilaterian-specific.

This analysis also included sequences identified from BLAST queries of bilaterian members which reportedly lack an ultrafiltratory capacity (Nematoda (=*C. elegans*) and Tunicata (=*C. intestinalis*). As discussed, both of these proteins carry out a diverse number of functions, and so the presence of orthologues for Neph and Nephrin in these taxa is perhaps not surprising. When these proteins adopted an ultrafiltratory function remains to be established, but it is clear that this role is conserved across a diverse number of bilaterian members, including representatives from the Deuterostomia, Lophotrochozoa and Ecdysozoa, and so a primitive

urbilaterian ultrafiltratory-specific function followed by loss in, for example *C. elegans* and *C. intestinalis*, is likely.

As shown in Appendix 5, bootstrap support at the nodes splitting off sequences for Neph1 and Nephrin is low. A similarly low bootstrap support value was returned even when putative Xenacoelomorpha orthologues were excluded from the alignment. A possible explanation for this is the high conservation of Ig-like domains between the two sequences. As both protein sequences were included for orthology inference in the same alignment, the high similarity of these domains could result in low support for the respective groupings of Neph1 and Nephrin. Despite this, protein domain order and structure of both Neph1 and Nephrin is identical between published vertebrate sequences and Xenacoelomorpha orthologues. Similarly, conserved domains (Ig-like in both proteins, and a fibronectin III-like repeat in Nephrin) taken from the vertebrate and Xenacoelomorpha sequences are well aligned. Taken together, these provide further evidence for the orthology of the Neph1 and Nephrin sequences identified in the Xenacoelomorpha.

4.2.2 Podocin orthologues and gene function

4.2.2.1 Presence and absence of Podocin-like sequences across the Eukaryota

Annotated orthologues of Podocin from NCBI are listed only from vertebrate taxa. In non-vertebrate taxa, orthologues of Podocin are named variably on NCBI as stomatin-like protein, erythrocyte band 7 (EB7) protein or Mechanosensory protein 2 (Mec2). Consequently, the best-hit sequence annotated as one of these proteins was included for analysis. Podocin-like sequences were identified in all diploblastic and triploblastic organisms (Figure 4.9). No potential orthologue was identified in the Choanoflagellatae. Outgroup sequences were chosen to represent proteins within a wider group classified by a 'SPFH' (Stomatin, Prohibitin, Flotillin, HflK/HflC) domain – that is, other protein families (specifically Prohibitins and Flotillins) with a domain that is similar to the stomatin-domain but specifically characteristic of these separate classes of proteins[157].

The maximum likelihood tree of these sequences grouped together all Podocin, Stomatin (=EB7) and Mec2 sequences from across the Metazoa (Figure 4.9). Sequences identified from Xenacoelomorpha transcriptomes group with all other bilaterian Podocin/EB7/Mec2 sequences. The outgroup sequences, chosen as Prohibitin and Flotillin, group together in their separate protein families, as would be expected. The long-branched stomatin-like protein sequences (SLP) cluster within the main Podocin/EB7/Mec2 grouping. In addition, the best-hit EB7 sequences identified from non-metazoan members (*C. owczarzaki* and *D. discoideum*) do not cluster with the main Stomatin group, but are instead found as a branch with EB7 sequences from a platyhelminth (*Echinococcus granulosus*), and the arthropods *Culex quinquefasciatus* and *Daphnia magna*. This finding contradicts a previous analysis that suggested a Podocin-orthologous protein is present outside of the Metazoa (Andrikou *et al*. preprint)[158]. Whilst we can assign the sequences identified from *C. owczarzaki* and *D. discoideum* as belonging to the wider Stomatin family, their placement indicates that they are not orthologous to metazoan Podocin/EB7/Mec2.

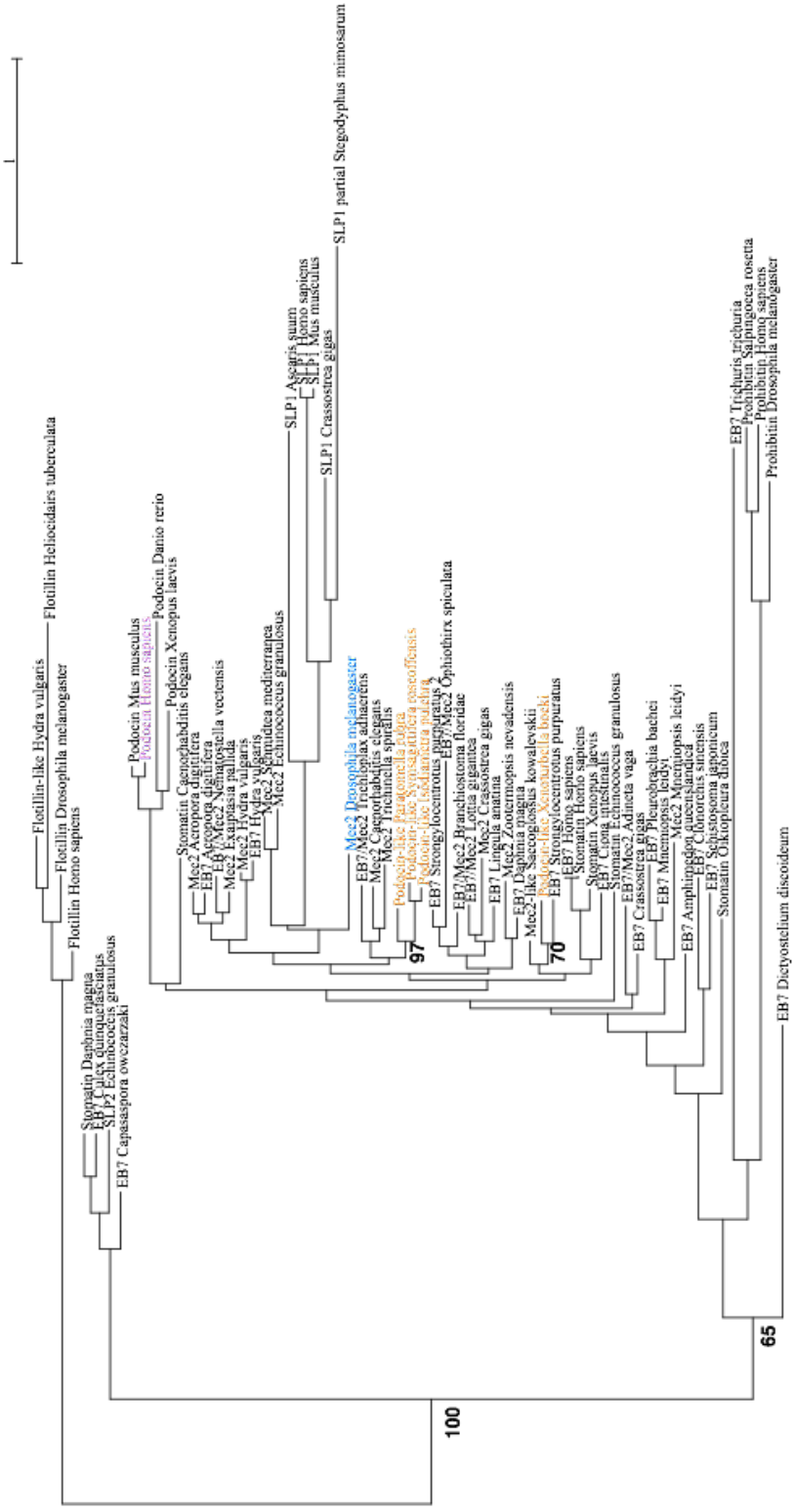


Figure 4.9. Maximum likelihood analysis of Podocin/EB7/Mec2 proteins and related outgroups. Xenacoelomorpha sequences shown in orange; *D. melanogaster* sequences in blue; *H. sapiens* sequences in purple. Phylogenetic inference carried out using RAxML. Bootstrap support values shown at relevant nodes. Branch length value shows number of substitutions per site.

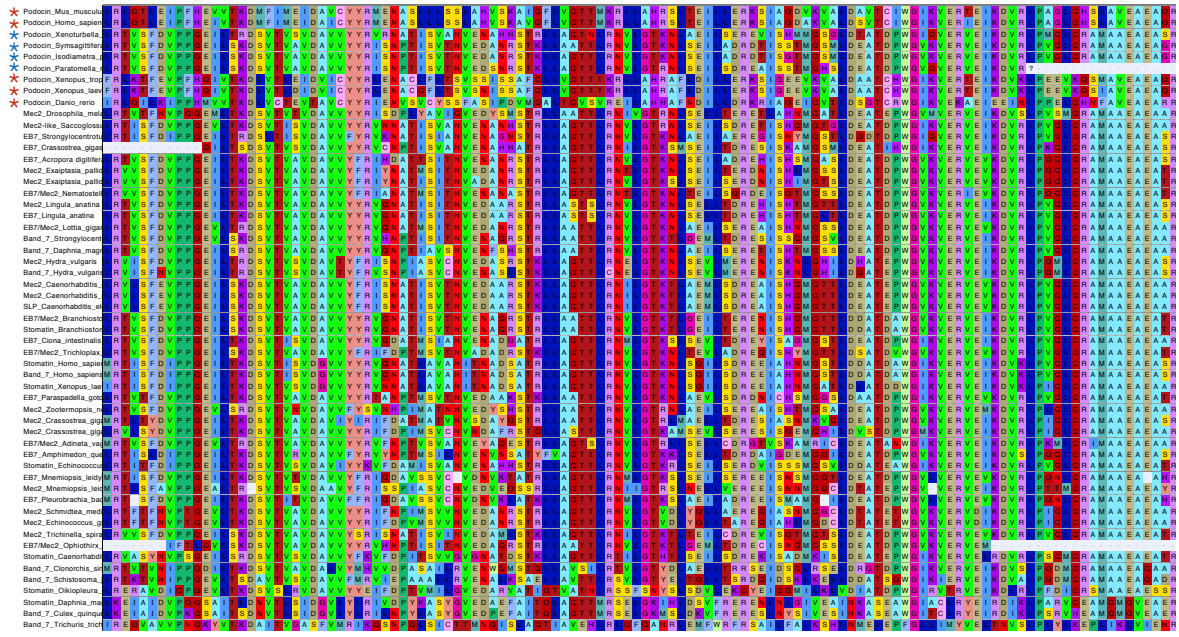
4.2.2.2 Function of Podocin and Stomatin-like proteins

Podocin belongs to the stomatin family of proteins, classified by the presence of a stereotypical and highly conserved region called the stomatin-domain[157] (Figure 4.10). Proteins with the stomatin-domain are numerous and found in all three classes of life, where they exhibit a remarkable degree of sequence conservation: homologs between bacteria and human have been found to share up to 50% amino acid similarity, and this degree of conservation is clear from the alignment of the stomatin-like domain in the diverse taxa included for phylogenetic analysis, including the Xenacoelomorpha (Figure 4.10)[159]. Phylogenetic analysis of the stomatin family of proteins suggests that eukaryotic stomatin proteins have two separate prokaryotic origins, and have subsequently undergone a series of duplication events to give the diversity of stomatin proteins found in different metazoan lineages. Stomatin-like-protein-3 (SLP-3/slipin-3) is purported to be vertebrate-specific, arising as a result of two gene duplication events, occurring before the teleost/tetrapod split[159]. Of the annotated vertebrate stomatin-domain proteins, Podocin is particularly 'long-branched', indicating a rapid rate of sequence evolution, and perhaps suggesting that the function of Podocin is quite different from the ancestral function of other stomatin proteins[159].

In vertebrates, Podocin (*NPHS2*) has kidney-specific expression, where it localises to the slit diaphragm of podocyte cells in the glomerular capillary wall[160]. Here, it co-localises with Nephrin and the actin cytoskeleton to act as a scaffolding protein and to maintain slit diaphragm integrity. Despite the kidney-specific function of Podocin, all mammalian stomatin-domain proteins share between 40-84% sequence identity at the level of the stomatin domain (Figure 4.10), and the unique functions of different stomatin proteins are thought to depend on the unique characteristics of the non-conserved N- and C- termini of the protein[157]. In *D. melanogaster*, the stomatin-domain protein Mec2 is said to be a homologue of mammalian Podocin. Much like Podocin expression in the vertebrate podocyte, Mec2 localises to nephrocytes in *D. melanogaster*, and is thought to maintain

integrity of the nephrocyte diaphragm and aid ultrafiltration[130]. Podocin/EB7/Mec2–like proteins are found throughout the Bilateria, where they also function in other roles outside of an ultrafiltratory capacity[161].

Figure 4.10. Alignment of the stomatin domain in Podocin/EB7/Mec2 sequences. Taxa used in phylogenetic inference, including Podocin-like orthologues identified in the Xenacoelomorpha. Vertebrate Podocin sequences are shown by a red asterisk; Xenacoelomorpha sequences are shown by a blue asterisk.



4.2.2.3 Podocin-like sequences in the Xenacoelomorpha

Sequences identified from Xenacoelomorpha transcriptomes group with all other bilaterian Podocin/EB7/Mec2 sequences (Figure 4.9). It is clear that the stomatin-like proteins have a very deep and conserved evolutionary history, as evidenced by the identification of related proteins in the Amoebozoa and Filasterea, and it is also evident that these proteins have a diverse function. Although Podocin is known to be vertebrate-specific, the expression of a stomatin-domain protein with similarity to Podocin in the *D. melanogaster* nephrocyte does suggest an orthologous function. Podocin/EB7/Mec2 orthologues are present throughout the Metazoa – as demonstrated by the inclusion of sequences in this analysis from diploblastic representatives – and so an ultrafiltratory-specific function is likely to have

arisen in the lineage leading to the Bilateria. Investigating the expression and function of this protein in the Xenacoelomorpha could therefore be informative for understanding its role more widely within bilaterian members.

4.2.3 CD2AP orthologues and gene function

4.2.3.1 Presence of CD2AP across the Eukaryota

The annotation of CD2AP/SH3-domain proteins is not consistent on NCBI, but single copies of CD2AP-like genes are known to be found across the Metazoa, and putative CD2AP sequences were identified in all bilaterians and diploblastic taxa investigated (Figure 4.11). Potential CD2AP-like sequences were identified in the Amoebozoa and Filasterea, but these are resolved with unclear orthology in phylogenetic analyses (Figure 4.11). Xenacoelomorpha sequences did not cluster as expected in phylogenetic analysis. Whilst sequences for *Xenoturbella* and *I. pulchra* are found in the main CD2AP grouping, the *P. rubra* sequence groups with the CDC25 outgroup sequences, and *S. roscoffensis* is found with the SH3-domain Intersectin outgroup (Figure 4.11).

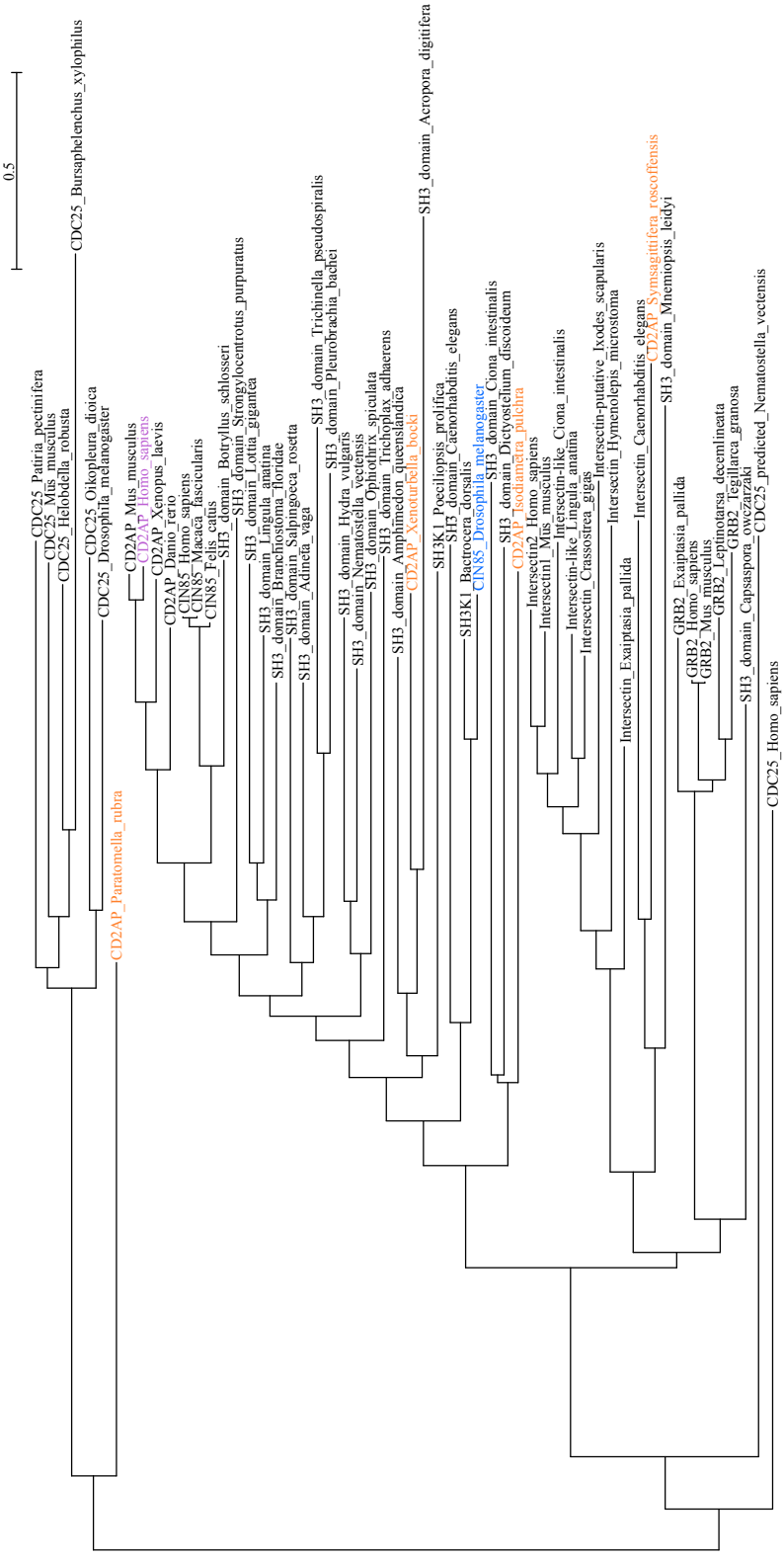


Figure 4.11. Maximum likelihood analysis of the SH3-domain CD2AP protein and related outgroups. Xenacoelomorpha sequences shown in orange; *D. melanogaster* sequences in blue; *H. sapiens* sequences in purple. Phylogenetic inference carried out using RAxML. Branch length value shows number of substitutions per site.

4.2.3.2 Structure of CD2AP

CD2AP is a multiple-SH3 domain protein, which, along with its paralogue CIN85, belongs to a protein family of adaptor molecules with roles in a number of cellular processes[162]. The presence of both CD2AP and CIN85 is unique to the vertebrates, and assumed to be the result of gene duplication after divergence from the invertebrates. Single copies of the gene are found in a number of other metazoan lineages[162]. In this analysis, CD2AP orthologues were identified in all bilaterian and diploblastic taxa investigated.

CD2AP is characterised by three SH3 domains, a proline-rich region, and a coiled coil region, and this conserved protein domain structure was found in all xenacoelomorph orthologues (Figure 4.12). SH3 domains are well conserved across all proteins in which they are found, and the conserved amino acid alignments of SH3-domain proteins – including those selected as outgroup sequences – make reliable phylogenetic inference more difficult.

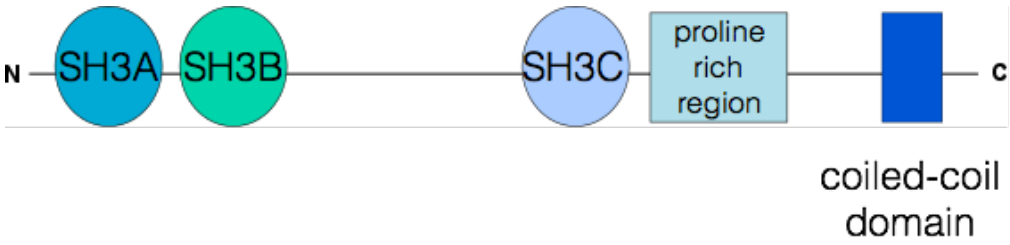


Figure 4.12. Schematic structure of CD2AP protein domains. These comprise three SH3 domains; a proline-rich region, and a coiled coil region at the C-terminus. Putative CD2AP orthologues identified in the Xenacoelomorpha have the same protein domains structure as is illustrated here.

SH3 domains taken from the Xenacoelomorpha sequences identified as putative orthologues of CD2AP were aligned with SH3 domains from the CD2AP sequences in *H. sapiens* and *M. musculus* (Figure 4.13). For all Acoela orthologues, transcriptome sequences appeared to be truncated – as such, no *P. rubra* SH3A domain and no *S. roscoffensis* or *I. pulchra* SH3C domains were included in alignment. It is evident from these alignments that SH3A and SH3B show the highest degree of conservation between *H. sapiens*, *M. musculus* and the Xenacoelomorpha sequences. However, the *Xenoturbella* sequence aligning to SH3B is very expanded between conserved blocks of residues, which appears to be unique compared to other SH3 domain proteins (Figure 4.13). It is possible that the difficulties in assigning orthology to the Xenacoelomorpha sequences could be a result of these sequence expansions and truncations. Both *Xenoturbella* and *I. pulchra* group with the other CD2AP sequences included for analysis in phylogenetic inference (Figure 4.11). Bootstrap support at the main CD2AP node is low (see Appendix 5), but the inclusion of other SH3-annotated protein sequences could have confounded confident phylogenetic inference. The conserved protein domain arrangement, and degree of alignment between SH3 domains taken from *Xenoturbella* and *I. pulchra* provide evidence for their orthology, but the unexpected placement of *P. rubra* and *S. roscoffensis* could be attributed to misidentification of the best-hit transcript from the CD2AP BLAST query.

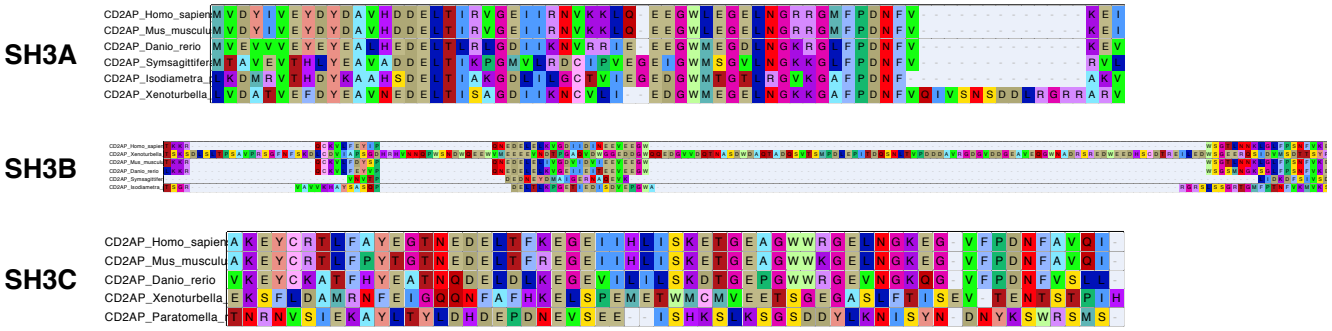


Figure 4.13. SH3A, SH3B and SH3C domains from putative Xenacoelomorpha CD2AP orthologues. Domains aligned to sequences taken from *H. sapiens* and *M. musculus* CD2AP.

4.2.3.3 Ancestral function of CD2AP

CD2AP is expressed in all vertebrate tissues, but its highest level of expression is in epithelial cells[141]. The primary purpose of proteins in the CD2AP/CIN85 family is proposed to be the downregulation of receptor tyrosine kinase activity during endocytosis, but it also functions in actin cytoskeleton regulation[163]. As described in 4.1.3, CD2AP is expressed in the podocytes of the vertebrate kidney, where it interacts with Nephrin to maintain the structure and function of the slit diaphragm[141].

The presence of CD2AP orthologues throughout the Metazoa means that this gene evolved prior to the evolution of ultrafiltratory specialisation within the Bilateria. CD2AP had broad expression in a number of different tissues in diverse taxa, where it is necessary for the binding of actin and cell adhesion molecules [163]. Thus is it likely that the interaction with Nephrin and Podocin at the slit diaphragm is a co-opted function.

4.2.4 ZO-1 orthologues and gene function

4.2.4.1 ZO- sequences in the Eukaryota

ZO-1 (Zona Occludens 1) is a tight junction protein belonging to the scaffolding membrane associated guanylate kinases (MAGUK) protein superfamily. ZO-1 orthologues were identified in all taxa investigated (Figure 4.14), with the exception of *D. discoideum*. Putative ZO-1 orthologues from the Xenacoelomorpha group with other ZO-1 proteins.

Cnidarians, the placozoan *Trichoplax adhaerans* and all bilaterian taxa are known to have at least one homolog from one of the seven main classes of MAGUK, including ZO. Whilst Porifera have a smaller MAGUK complement and lack a reliable homologue for a ZO protein, they do have protein sequences resembling that of the larger disc-large (DLG) super class of proteins (which is part of the MAGUK family) but without clear assignment to any of the classes therein[164]. Similarly, both the choanoflagellates and filastereates have proteins belonging to the DLG class, but these cannot be confidently identified as ZO homologues. The 'best-hit' ZO sequence for *C. owczarzaki* found in this analysis places it with the other MAGUK outgroups, and not within the node representing ZO- sequences (Figure 4.14). No sequences resembling the SLG class could be identified in *D. discoideum*, representing the Amoebozoa, which supports the proposed theory that the SLG super class originated in the Holozoa[164].

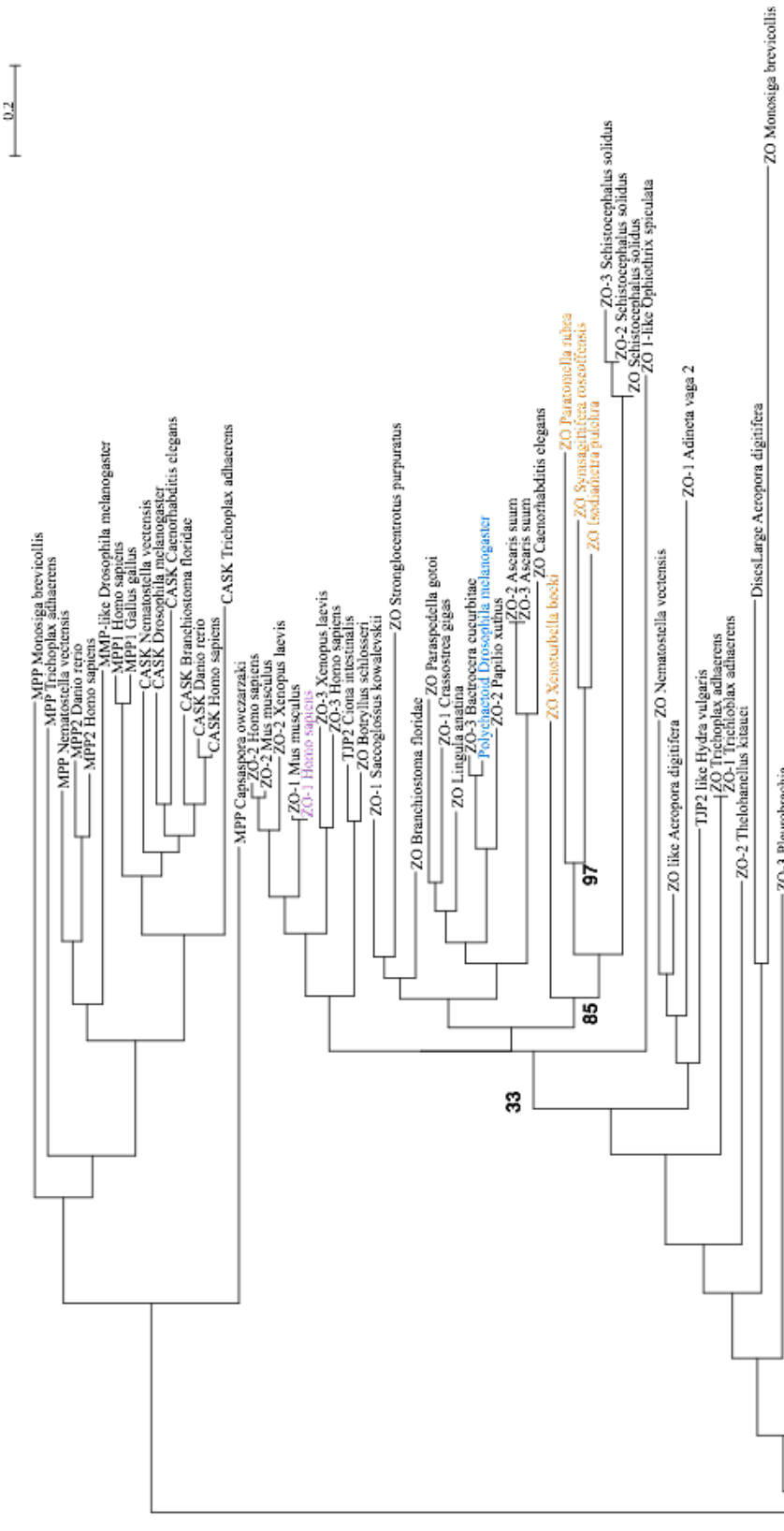


Figure 4.14. Maximum likelihood analysis of the tight junction protein ZO-1 and related outgroups. Xenacoelomorpha sequences shown in orange; *D. melanogaster* sequences in blue; *H. sapiens* sequences in purple. Phylogenetic inference carried out using RAxML. Bootstrap support values shown at relevant nodes. Branch length value shows number of substitutions per site.

4.2.4.2 ZO- protein structure

Proteins in the MAGUK superfamily are defined by an inclusion of PDZ, SH3 and Guanylate Kinase (GuK domains) (Figure 4.15A)[165]. Aligning the three PDZ domains and GuK domain from *Xenoturbella*, *P. rubra*, *S. roscoffensis* and *I. pulchra* against the corresponding *H. sapiens* and *M. musculus* domains shows a high level of conservation at these domains between taxa, but with some sequence expansion between conserved residues in the Xenacoelomorpha GuK domains (Figure 4.15B).

4.2.4.3 Ancestral function of ZO- proteins

The MAGUK class of proteins is evolutionary ancient, and can be broadly divided into eight monophyletic groups, plus one paraphyletic group, found throughout the Eukaryota[164]. They have a number of intercellular roles: ZO proteins function broadly as scaffolding proteins at cellular junctions, and also regulate cell growth and proliferation. As described in 4.1.3, ZO-1 expression is necessary in vertebrates for establishing the podocyte filtration barrier[166], and the expression of a ZO-1 orthologue in the nephrocytes of *D. melanogaster* is also required to mediate ultrafiltration[130].

Nonetheless, it is clear that ZO proteins, including ZO-1, have a broad expression pattern and carry out diverse functions associated with cell growth, proliferation, scaffolding, and intracellular signalling[167]. Whilst ZO-1 plays a critical role in establishing the filtration barrier in podocytes and nephrocytes, the presence of this protein in non-bilaterian taxa, and its broad expression and function, means that this filtratory-specific expression was likely co-opted specifically within the Bilateria. Consequently, the identification of orthologous ZO-1 protein sequences in the Xenacoelomorpha does not imply an ultrafiltratory capacity.

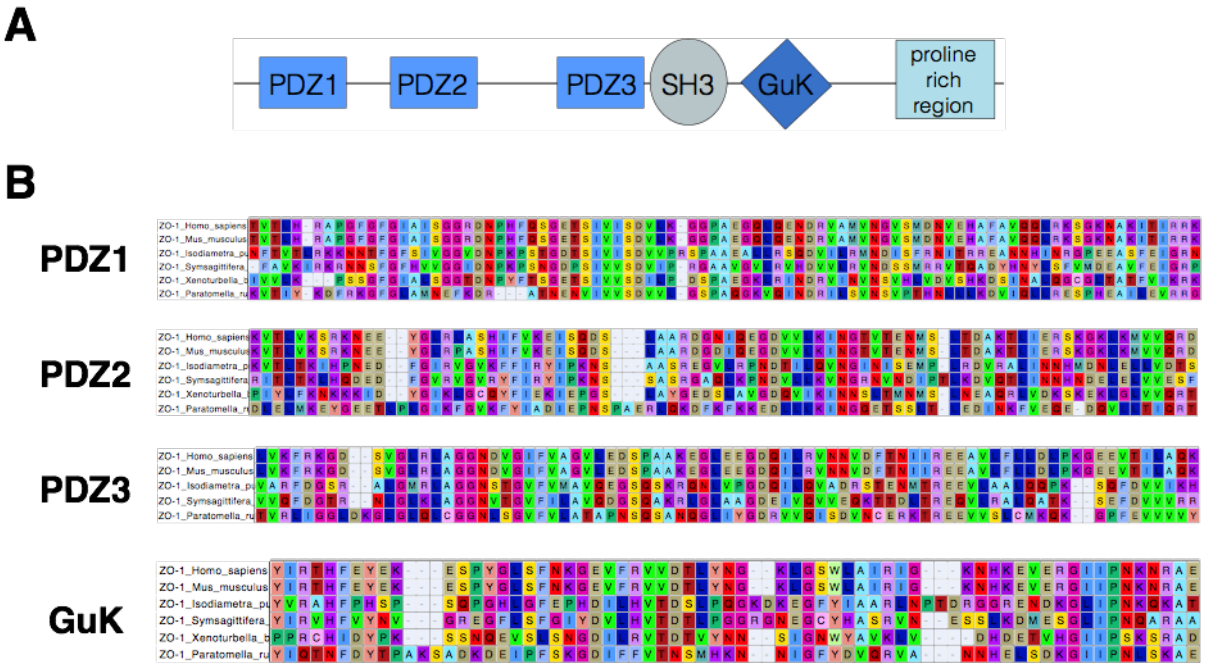


Figure 4.15. Structure of ZO-1 conserved protein domains. (A) Schematic structure of the ZO-1 protein, comprising three PDZ domains, one SH3 domain and a Guanylate Kinase (GuK) domain, with a proline rich region at the C terminus. (B) PDZ1, PDZ2, PDZ3 and GuK domains from putative Xenacoelomorpha ZO-1 orthologues aligned against the corresponding domains in *H. sapiens* and *M. musculus*.

4.3 General conclusions

An overview of the presence and absence of ultrafiltratory-related genes in the Eukaryota as a result of the analysis described in the chapter is presented in Figure 4.16.

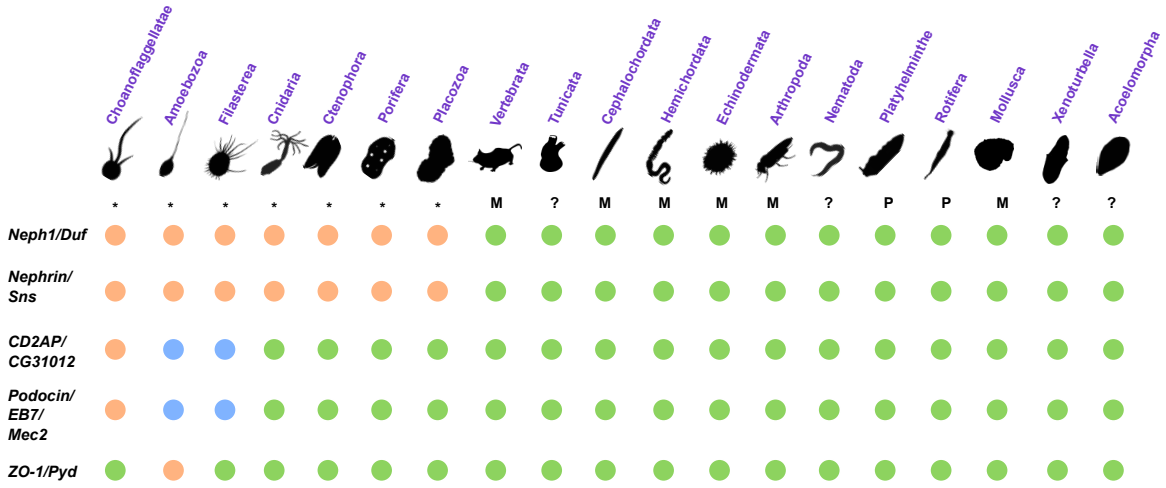


Figure 4.16. Presence and absence of ultrafiltratory-related genes in the Metazoa. Presence of orthologues indicated by green circles; genes of unclear orthology indicated by blue circles; absence indicated by orange circles. Transcriptome and genome mining in: *Monosiga brevicollis* (Choanoflagellatae); *Dictyostelium discoideum* (Amoebozoa); *Capsaspora owczarzaki* (Filasterea); *Acropora digitifera*, *Nematostella vectensis*, *Hydra vulgaris* (Cnidaria); *Mnemiopsis leidyi*, *Pleurobrachia bachei* (Ctenophora); *Amphimedon queenslandica* (Porifera); *Trichoplax adhaerans* (Placozoa); *Homo sapiens*, *Mus musculus*, *Danio rerio*, *Xenopus laevis* (Vertebrata); *Botryllus schlosseri*, *Ciona intestinalis* (Tunicata); *Branchiostoma floridae* (Cephalochordata); *Saccoglossus kowalevskii* (Hemichordata); *Ophiothrix spiculata*, *Strongylocentrotus purpuratus* (Echinodermata); *D. melanogaster* (Arthropoda); *Caenorhabditis elegans* (Nematoda); *Echinococcus granulosus*, *Schmidtea mediterranea* (Platyhelminth); *Adineta vaga* (Rotifera); *Lottia gigantea*, *Crassostrea gigas* (Mollusca); *Xenoturbella bocki*, *Paratomella rubra*, *Symsagittifera roscoffensis*, *Isodiametra pulchra* (Xenacoelomorpha). Type of nephridial system indicated by: M = metanephridia; P = protonephridia; ? = unclear or presumed absence; * = no nephridial or excretory system.

From transcriptome-mining and phylogenetic inference, it is clear that all five ultrafiltratory-related proteins investigated have orthologues within the Xenacoelomorpha. For Podocin/EB7/Mec2, CD2AP and ZO-1, the presence of these orthologues is expected, given their existence throughout metazoan and/or wider eukaryotic taxa. Although these proteins have been identified at the site of ultrafiltration in vertebrates and *D. melanogaster*, they also function widely in other diverse roles, and so their presence in the Xenacoelomorpha is not necessarily indicative of any ultrafiltratory capacity. However, the identification of Neph1 and Nephrin in the Xenacoelomorpha is perhaps a more interesting find, given their absence in any organisms outside of the Bilateria. The presence of Xenacoelomorpha orthologues of these sequences thus provides the opportunity for molecular investigation into the expression and function of these genes in a phylum that is thought to lack any ultrafiltratory capacity. Furthermore, the evolutionary origin and conservation of ultrafiltratory and excretory systems remains inconclusive. The intriguing phylogenetic position of the Xenacoelomorpha means that investigating the role of these genes – and of Neph1 and Nephrin in particular – could perhaps be informative for our understanding of the degree of conservation of these genes with regards to their role in ultrafiltration in the Bilateria.

5 Molecular approaches in *Symsagittifera roscoffensis*

5.1 Introduction

5.1.1 Ultrafiltration and excretory systems in the Acoela

As is true for all acoels, *Symsagittifera roscoffensis* is a small marine worm with a simple body plan. One aspect of the simple morphology of acoels, and indeed, all taxa within the Xenacoelomorpha, is that they are commonly regarded to lack any structures relating to ultrafiltration or excretion (see section 1.4.1). Two alternate theories pervade regarding the placement of the Xenacoelomorpha within the Bilateria: either as sister group to the Ambulacraria within the deuterostomes, or at the base of the Bilateria, separate from the main protostome and deuterostome grouping. Reflecting this lack of nephrocytes, the proponents of positioning the xenacoelomorphs as an early branch refer to the protostome/deuterostome clade as the Nephrozoa: the absence of nephrocyte-like structures is seen as a significant criterion for the exclusion of the Xenacoelomorpha from this grouping. I have shown in Chapter 4 that the xenacoelomorphs possess five of the genes that are co-expressed at the site of ultrafiltration in other bilaterians. In this chapter, I aim to use *in situ* hybridisation and antibody staining to establish the expression domains of purported ultrafiltratory-related genes in the acoel *S. roscoffensis*.

5.1.2 Selecting molecular markers of ultrafiltration

As discussed in 4.1.3 many lines of evidence suggest that the expression of five structural proteins (Neph1, Nephrin, Podocin, ZO-1, and CD2AP) is conserved at the site of ultrafiltration in the three main bilaterian lineages: Deuterostomia (investigated in vertebrates)[136]; Ecdysozoa (*D. melanogaster*)[130]; and Lophotrochozoa (*S. mediterranea*)[37]. Of these

proteins, it is the co-expression of Neph1 and Nephrin to form heterodimers and homodimers (specifically Nephrin-Nephrin) that are critical for the formation and function of the ultrafiltratory diaphragm. Although these proteins function more widely in a number of different processes, the co-localisation of these proteins in the same cell is likely to be a marker of ultrafiltration.

Podocin is a protein that is found exclusively in the vertebrate podocyte diaphragm. Although assigning orthology to Podocin/EB7/Mec2 sequences found outside of the vertebrates is not straightforward, expression and function of a Podocin-like sequence (=EB7/Mec2) in the nephrocytes of *D. melanogaster* suggests that the ultrafiltratory function of this protein is conserved in the Bilateria[130]. Given the uncertain position of xenacoelomorphs within the Bilateria, investigating the expression of transcribed gene sequences identified as potential orthologues of Podocin – and whether they co-localise with Neph1 and/or Nephrin – may be informative for our understanding of the historic function of this gene.

Although CD2AP and ZO-1 are also known to have a role in ultrafiltration, they also function much more broadly in a number of different roles, and so the primary focus of this analysis was on understanding the expression pattern of Podocin, Neph1 and Nephrin.

5.1.3 *Symsagittifera roscoffensis*: morphology and use for molecular lab approaches

As we have seen in section 4.1.3**,** BLAST searches of members of the Acoelomorpha (*S. roscoffensis, I. pulchra, and P. rubra*) identified putatively orthologous sequences of these three genes-of-interest. *P. rubra* specimens can be collected from sand in the inter-tidal zone in Yorkshire. Very little is known about their habitat and ecology and *P. rubra* has not been established in culture in any lab to date, making it a poor choice for any long-term molecular lab work. *I. pulchra*, along with *H. miamia*, is as close to a 'model'

acoel as is currently established. It can be maintained in culture with varying degrees of success, but is prone to population crashes and severe ciliate contamination, which results in reduced egg laying and a rapid decline in population health.

Although *S. roscoffensis* cannot be maintained long-term in culture, it has previously been used for experimental lab work[168-170], and offers an alternative experimental acoel species to *I. pulchra*, without the problems associated with maintaining cultures. *S. roscoffensis* lives abundantly in huge colonies in the intertidal regions across Northern European coasts, and can be easily collected in significant quantities at low tide (Figure 5.1). Adults are bright green owing to their endosymbiotic relationship with the algal species *Tetraselmis convolutae*, and once in the lab, can be kept in incubators on a 12-hour light/dark cycle for a number of weeks[171]. Adults lay numerous eggs in the first few days following collection in the breeding season of September to June: up to 30 eggs are laid per individual, enclosed as multiple eggs within a single cocoon. In the lab, embryonic development takes up to four days, at which point animals hatch as colourless, free-living juveniles. Without the presence of the symbiont, juveniles do not live longer than ~20 days, but can be fixed as healthy individuals at different developmental stages prior to this point.

S. roscoffensis is a hermaphroditic, soft-bodied animal (Figure 5.1). Juvenile animals up to approximately seven days old grow up to 2mm in length and are teardrop shaped, with a broad anterior portion narrowing towards the posterior. Adults are up to 4mm in length, and more vermiform, losing the teardrop shape found in juveniles. The body wall comprises multi-ciliated epidermal cells interspersed with muscle cells and gland openings of 1μm diameter[172]. The mouth is found in the centre of the body axis in juvenile animals, and in the anterior third of the body in adults. Anterior to the mouth, both adult and juvenile animals are grooved along the longitudinal axis, formed by the ventral bending of the lateral edges of the animal: in juveniles this gives animals a cup-like appearance. The male gonopore is found at the very posterior of the animal, and the female gonopore is found at the

anterior-most third[172]. A synapomorphy of the taxon Sagittiferidae – to which *S. roscoffensis* belongs - is the presence of sagittocysts – capsules containing a protrusible needle-like structure, thought to function with the male and female copulatory organs.

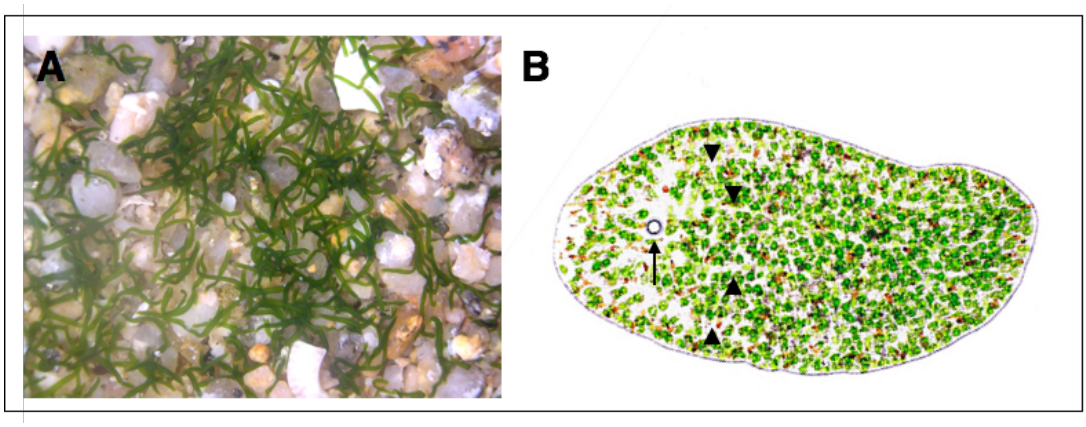


Figure 5.1. The ecology and morphology of *Symsagittifera roscoffensis*. (A) photograph taken from http://biologie.ens-lyon.fr/ressources/Biodiversite/Documents/image-de-la-semaine/2011/semaine-46-14-11-2011 showing an abundance of *S. roscoffensis* living in the intertidal region. (B) the simple morphology of the adult *S. roscoffensis*, anterior to the left. The animal is green owing to the symbiotic relationship with *Tetraselmis convolutae*; navigational statocyst indicated by black arrow; arrowheads showing muscle striation.

Some gene visualisation protocols have been published for *S. roscoffensis*, and techniques for immunohistochemistry and *in situ* hybridisation have been successful in informing our knowledge of the acoel nervous system and the expression of various muscle genes[168-170]. Nevertheless, these protocols have not been consistently successful: published gene expression studies are largely limited to juvenile animals (up to one week old), and the symbiotic relationship with *T. convolutae* poses a problem for background auto-fluorescence in adult immunohistochemistry experiments with fluorescent signal. Consequently, establishing and fine-tuning experimental procedures in *S. roscoffensis* is an ongoing challenge.

5.1.4 Objectives of Chapter

The objective of the work described in this chapter was to look for evidence of nephrocytes in the Acoelomorpha by investigating the expression patterns of genes with a role in ultrafiltration, using *S. roscoffensis* as an experimental species. It is known from studies in other animals that the five core ultrafiltratory related genes are expressed and function outside of the ultrafiltratory apparatus, but looking for cells in which these genes are co-expressed would likely be a marker of ultrafiltratory capacity in a taxon that is widely believed to lack any specialised filtratory or excretory system. Sequences orthologous to *Neph1*, *Nephrin* and *Podocin* were identified in the transcriptome of *S. roscoffensis* and will subsequently be referred to as *SrNeph1*, *SrNephrin* and *SrPodocin-like*.

The acoels remain a difficult group to investigate experimentally, and *S. roscoffensis* is not well established as a model taxon. Consequently, a significant objective of this chapter was to develop and troubleshoot gene visualisation protocols in this animal at different developmental stages. Establishing more reliable experimental protocols for this animal will be beneficial not only for the objectives of this chapter, but also for wider investigations into the morphology and genetics of this species.

5.2 Results and Discussion

Owing to the difficulties encountered with gene visualisation approaches in *S. roscoffensis*, results and discussion are presented together in this chapter as a compound section. Whilst some details of experimental protocols are outlined below, detailed methods are described in sections 2.8 and 2.9

5.2.1 *In situ* hybridisation in adult *S. roscoffensis*

5.2.1.1 SrTroponin I positive control

In order to establish a working *in situ* hybridisation protocol in *S. roscoffensis*, I designed a probe for *SrTroponin I*, a commonly used muscle marker with known domains of expression in the Acoela. Using the PBS-based protocol described by Chiodin *et al.* (2011)[170] (based on the original protocol published by Semmler *et al*. (2010)[168]), I was able to recapitulate the expression of *SrTroponin I* in whole-mount and sectioned adult animals (Figure 5.2). *SrTroponin I* in adults is expressed in the posterior part of the animal; in the male gonopore; in a longitudinal band anterior to it through the midline of the animal; and in the sagittocysts at the posterior tip and two longitudinal bands near to the epidermis along the left and right sides of the animal (Figures 5.2A and 5.2B). This expression of *SrTroponin I* was observed in all *in situ* hybridisation protocols using both whole mount and sectioned adult *S. roscoffensis*.

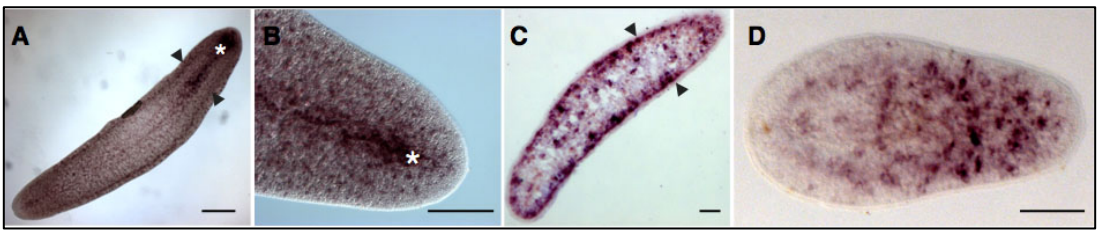


Figure 5.2. Control *in situ* hybridisation experiments for *SrTroponin I* in *S. roscoffensis.* Anterior is to the left in all aspects. (A) *SrTroponin I* in whole-mount adult using a PBS-based *in situ* hybridisation protocol. Expression is highest in the male gonopore at the posterior (asterisk), in a longitudinal band directly behind it, and in two bands down the length of the animal (arrow heads). (B) Detail of *SrTroponin I* expression in the male gonopore, from a PBS-based *in situ* hybridisation protocol (asterisk). (C) *SrTroponin I* expression in sectioned adult animal, using the protocol established for sectioned *Xenoturbella*: expression recapitulates that observed in (A). (D) *SrTroponin I* expression in whole-mount juvenile using a MABT-based protocol. Expression is scattered, with highest region of expression in the posterior of the animal. Scale bars in A.B,C: 100μm; D: 50μm.

5.2.1.2 SrNeph1 and SrNephrin expression in adult S. roscoffensis

Having succeeded with the control PBS-based *in situ* hybridisation protocol in adult *S. roscoffensis*, I used the same approach to investigate expression of the ultrafiltratory genes-of-interest. Probes were synthesised as outlined in sections 2.6.3 - 2.6.5. For probe length and approximate location in the complete sequence see Appendix 4. Initial repeats of the protocol using RNA probes for *SrNeph1* and *SrNephrin* gave inconclusive results in *S. roscoffensis* adults, and a number of experimental variations were therefore tried to troubleshoot the protocol. These included varying probe concentration from 0.01ng/μl to 10ng/μl; varying hybridisation temperature from 65°C (with the hope of high specificity reducing background staining) and at various lower temperatures down to 50°C (with the hope of lower stringency increasing the likelihood of signal); varying duration of hybridisation (from overnight to one week); and varying the concentration of NBT/BCIP in AP buffer used in the final signal development stage (see section 1.2). Nonetheless, results remained inconsistent between different rounds of experiments, with either no discernible signal at all; the same

expression pattern for all three genes and corresponding sense probes; or high background expression.

Of all rounds of *in situ* hybridisation using the PBS-based protocol, a possible expression pattern for *SrNeph1* and *SrNephrin* was achieved using probes at a concentration of 2ng/µl and incubating for three days at 60°C (Figure 5.3). Both genes appear to be expressed in a feathered expression in longitudinal bands either side of the midline of the animal. For *SrNeph1*, expression is constrained more closely to the midline, with much more enhanced staining in the posterior region of the animal (Figure 5.3A). A similar pattern is seen for *SrNephrin*, but with more broad bands of expression extending further away from the midline of the animal (Figure 5.3B). However, the difficulties associated with implementing *in situ* hybridisation for these probes means that I am not confident in the validity of these domains of expression for *SrNeph1* and *SrNephrin*.

Figure 5.3. Expression of *SrNeph1* and *SrNephrin* in whole-mount adult *S. roscoffensis* using a PBS-based *in situ* hybridisation protocol. Anterior is to the left in all aspects. (A) *SrNeph1* (B) *SrNephrin*. Expression of both genes appears to be in longitudinal feathered bands either side of the gut. Scale bars in all aspects: 100μm.

Changing *in situ* protocols entirely, I used a maleic acid based method to see if an alternative buffer might generate better results. Whilst this approach worked well for the *SrTroponin I* control probe, the ultrafiltratory gene RNA probes gave a great deal of dark NBT/BCIP background staining, with little difference between sense and anti-sense probes, even when the hybridisation temperature was increased with the aim of increased hybridisation stringency (Figure 5.4). A possible explanation for the ubiquitous staining across the animal is the abridged preparatory steps used in the maleic acid *in situ* protocol. In the PBS-based protocol I used triethanolamine and acetic anhydride washes as an acetylation step to reduce non-specific binding and therefore reduce NBT/BCIP background staining during probe visualisation. In the maleic acid protocol, specimens were washed straight from rehydration gradient MeOH:DEPC-H_2O washes into maleic acid washes into hybridisation buffer, with no additional steps to

reduce non-specific binding. This could have caused the dark background staining which developed very quickly (within 30 minutes) across the animal. In addition, this protocol omitted digestion with Proteinase K. Adult *S. roscoffensis* have a ciliated epidermis scattered with gland cells: brief treatment with Proteinase K is thought to be sufficient to permeabilise the epidermis for probe binding, but without causing over-digestion. The low Tween-20 concentration in the MABT solution, and the absence of a Proteinase K step in this protocol, could have prevented penetration of the probe, causing background staining across the epidermis of the animal.

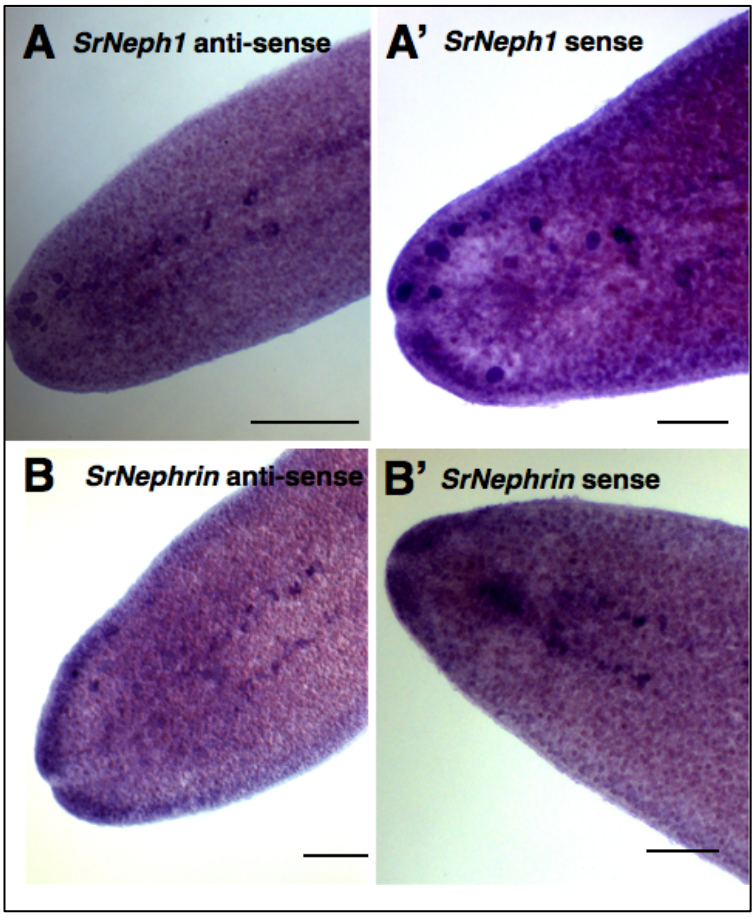


Figure 5.4. Expression of *SrNeph1* and *SrNephrin* in whole-mount adult *S. roscoffensis* using a MABT-based *in situ* hybridisation protocol. Anterior is to the left in all aspects. (A) *SrNeph1* anti-sense probe; (A') *SrNeph1* sense probe; (B) *SrNephrin* anti-sense probe; (B') *SrNephrin* sense probe. All probes show dark background staining and enhanced expression in longitudinal bands running the length of the animal, indicative of non-specific binding. Scale bars in all aspects: 100µm.

5.2.1.3 SrPodocin-like expression in adult S. roscoffensis

Using the PBS-based protocol for *SrPodocin-like* revealed less broad domains of expression, with signal localised to channels extending perpendicular to the length of the animal (Figure 5.5). This pattern is reminiscent of that seen for *SrNeph1* and *SrNephrin*. Nonetheless, given the difficulties in revealing any *in situ* hybridisation signal using this protocol, I am again skeptical of the validity of this result.

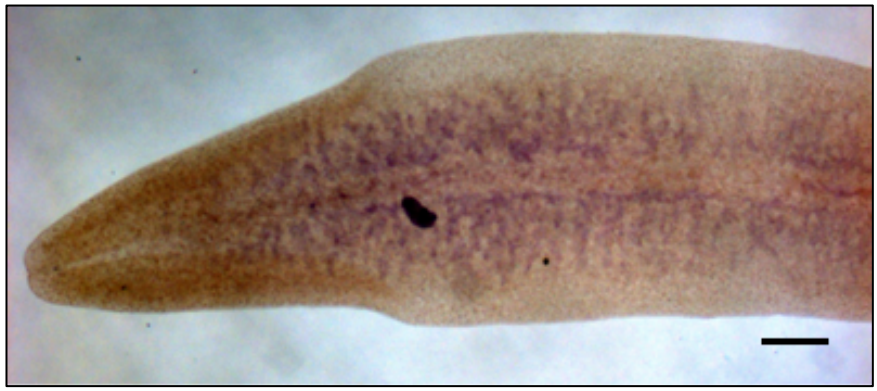


Figure 5.5. Expression of *SrPodocin-like* in whole-mount adult *S. roscoffensis* using a PBS-based *in situ* hybridisation protocol. Anterior is to the left. Expression appears to be in longitudinal feathered bands either side of the gut and is reminiscent of that seen for *SrNeph1* and *SrNephrin*. Scale bar: 100μm.

Using the same MABT protocol as used for *SrNeph1* and *SrNephrin* also resulted in ubiquitous background staining for both the *SrPodocin-like* sense and anti-sense probe (Figure 5.6).

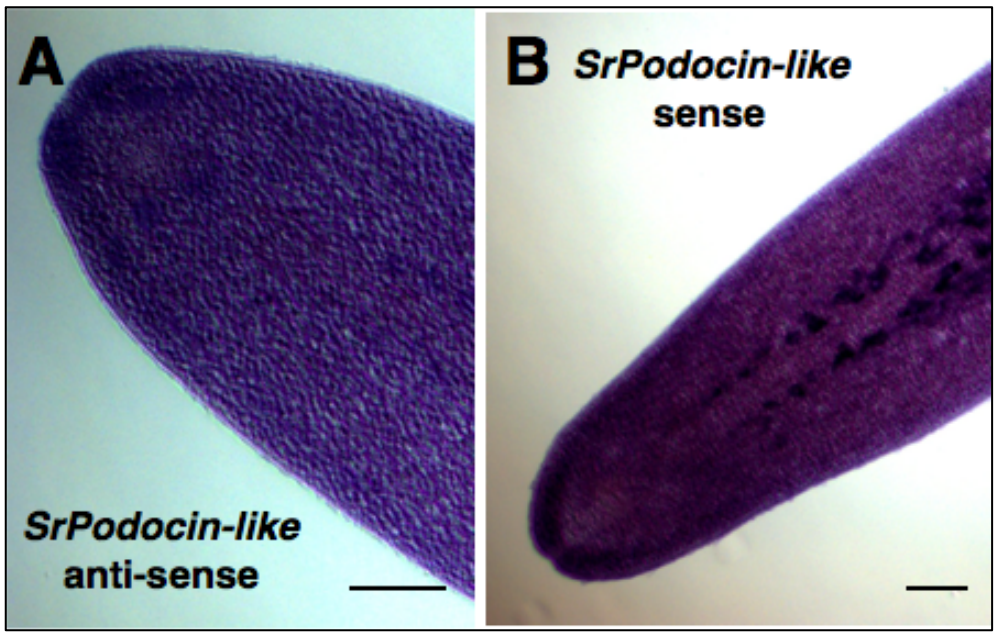


Figure 5.6. Expression of *SrPodocin-like* in whole-mount adult *S. roscoffensis* using a MABT-based *in situ* hybridisation protocol. Anterior is to the left in all aspects. (A) *SrPodocin-like* anti-sense probe (B) *SrPodocin-like* sense probe. As seen for *SrNeph1* and *SrNephrin*, both probes show dark background staining and expression in longitudinal stripes running the length of the animal, indicative of non-specific binding. Scale bars in all aspects: 100µm.

Finally, I used the *in situ* protocol developed for sectioned *Xenoturbella* specimens (see 2.8.1), on sectioned adult *S. roscoffensis*, with the hope that this might aid probe penetration without the risk of over-digestion. This gave a reliable result for the control *SrTroponin I* probe (Figure 5.2C), but no obvious signal could be seen for *SrNeph1* or *SrNephrin*. Successive sections appear to show a signal for *SrPodocin-like* at the anterior tip of the animal and in two lateral stripes running parallel to the midline of the animal (Figure 5.7). Given the absence of background staining, and the consistent domains of expression for *SrPodocin-like* seen moving through progressive sections of the animal, I suspect that this result is more likely to be a true expression pattern compared to those observed using different protocols.

Figure 5.7. Expression of *SrPodocin-like* in sectioned adult *S. roscoffensis*. From left to right sections move from close to the epidermis (A), through to the gut (A''). Anterior to the left. Expression localises to two longitudinal bands of cells flanking the gut, with the strongest expression seen in A'. Scale bar: 100μm.

5.2.2 *In situ* hybridisation in juvenile *S. roscoffensis*

After the largely inconclusive and/or unsuccessful outcomes for the ultrafiltratory genes-of-interest in adult animals, I repeated *in situ* hybridisation using both protocols for *SrTroponin I*, *SrNeph1*, *SrNephrin* and *SrPodocin-like* on fixed juvenile (3 to 7 day-old) animals.

In whole-mount juvenile animals, the control probe for *SrTroponin I* successfully recapitulated results described by Chiodin *et al.* (2011)[170]. Expression is widespread throughout the juvenile animal, but enhanced at the posterior end and in the mouth region in the centre of the animal (Figure 5.2D).

5.2.2.1 SrNeph1 expression in juvenile S. roscoffensis

SrNeph1 expression in juvenile animals appeared to be consistent in both *in situ* hybridisation protocols (Figure 5.8A and 5.8B). More background staining is evident in the original PBS protocol. *SrNeph1* appears to be expressed in two thin lateral stripes in the peripheral parenchyma on either side of the central gut, starting posteriorly to the statocyst and continuing until the last third of the body. At the start of these two bands is a strong region of expression running perpendicular to the two bands to join them across the

midsection of the animal. Expression is also found in the anterior region of the animal, with a degree of low-level expression appearing as a web across the epidermis of the animal, although this is more enhanced in the PBS-based protocol.

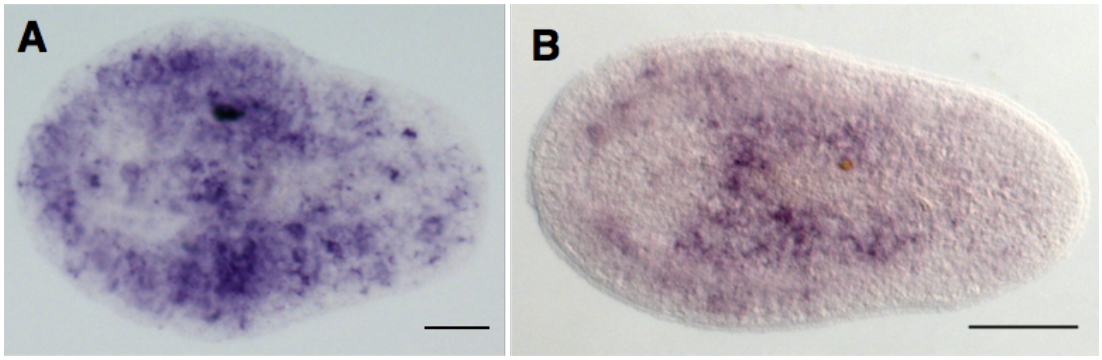


Figure 5.8. Expression of *SrNeph1* in whole-mount juvenile *S. roscoffensis*. Anterior is to the left in all aspects. (A) *SrNeph1* PBS protocol; (B) *SrNeph1* MABT protocol. Expression using both protocols identified in the parenchyma flanking the gut and in a lateral band across the midline of the animal, posterior to the statocyst. Scale bars in A and B: 50μm.

5.2.2.2 SrNephrin expression in juvenile S. roscoffensis

SrNephrin results are less consistent, but appear to show expression across the epidermis of the animal (Figure 5.9A and 5.9B). In addition, expression is enhanced in the centre of the juvenile, although signal is more ubiquitous than that observed for *SrNeph1,* especially using the MABT-based protocol.

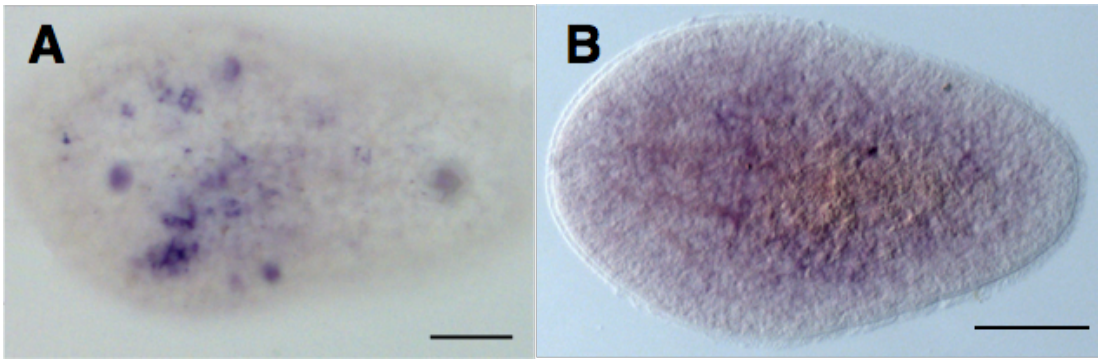


Figure 5.9. Expression of *SrNephrin* in whole-mount juvenile *S. roscoffensis*. Anterior is to the left in all aspects. (A) *SrNephrin* PBS protocol; (B) *SrNephrin* MABT protocol. The MABT protocol shows a lot of unspecific background staining, but expression appears to be enhanced in the centre of the juvenile. Scale bars in A and B: 50µm.

5.2.2.3 SrPodocin-like expression in S. roscoffensis

SrPodocin-like expression is also found across the epidermis of the animal, but with strongest expression in the centre of the animal, in the anterior portion of the ventral groove, and in the statocyst and the region directly posterior to the statocyst (Figures 5.10A and 5.10B). In the MABT protocol, *SrPodocin-like* shows a lot of background staining: expression domains for this probe are less consistent than those for *SrNeph1* and *SrNephrin*. Interestingly, *SrPodocin-like* is always detected very strongly in the statocyst (Figure 5.10).

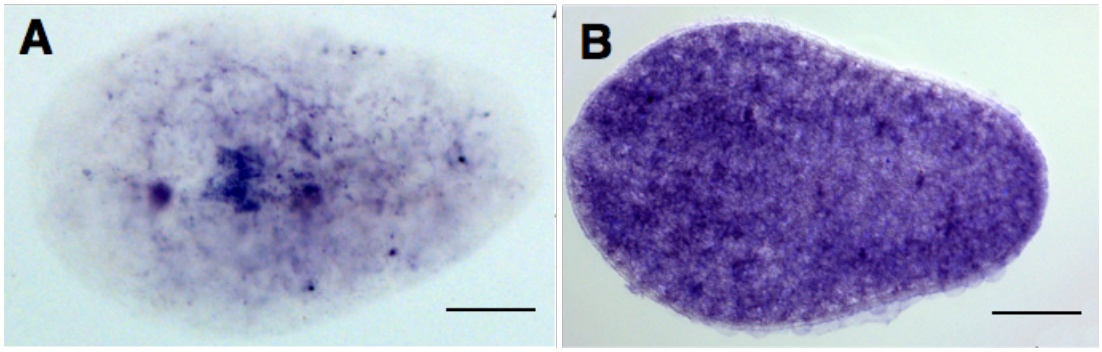


Figure 5.10. Expression of *SrPodocin-like* in whole-mount juvenile *S. roscoffensis*. Anterior is to the left in all aspects. (A) *SrPodocin-like* PBS protocol; (B) *SrPodocin-like* MABT protocol, hindered by a high level of ubiquitous background staining. Scale bars in A and B: 50μm.

in situ hybridisation experiments using juvenile *S. roscoffensis* revealed more consistent expression patterns than those found in the adult. *SrNeph1* is found in the parenchymal region; *SrNephrin* and *SrPodocin-like* show a more ubiquitous expression, but both probes show strongest expression in the central gut region of the adult. These results suggest broad domains of expression for these ultrafiltratory genes in the juvenile animal, but with strongest domains of expression in or close to the central syncytial gut.

5.2.3 Immunohistochemistry: commercial antibody expression in *S. roscoffensis*

5.2.3.1 Selection of commercial antibodies for use in S. roscoffensis

As an alternative visualisation approach, commercial antibodies raised against vertebrate for Nephrin, Neph1 and Podocin were used in adult and juvenile specimens. Numerous commercial antibodies against Neph1 (anti-Kirrel), Nephrin (anti-NPHS1) and Podocin (anti-NPHS2) are available, most commonly designed to investigate renal disease in humans. The antigens of these antibodies are against the human proteins. Selection of an appropriate antibody was based on two criteria:

1. For each antibody (against Neph1, Nephrin or Podocin), the published antigen sequence from each company was aligned to the translated sequences identified from the *S. roscoffensis* transcriptome for the appropriate gene. The antigen sequence that had the greatest similarity to that of *S. roscoffensis* was, where possible, chosen as a possible antibody.

2. As co-localisation of the proteins was an important consideration for investigating gene expression, the host animal for antibody production for Nephrin needed to be different from the host used for Neph1 and Podocin. As the vast majority of commercial antibodies against these genes are raised in rabbit, very few alternative options were available for Nephrin. Only one appropriate antibody, raised in sheep, was identified for this gene.

All commercial antibodies tried were polyclonal, giving them the possible advantage of being able to identify homologous proteins outside of the species from which the antigen was taken. Nonetheless, it is acknowledged that despite using polyclonal antibodies, similarity between the human antigen and the *S. roscoffensis* protein-of-interest may not be high enough for any protein detection. Similarly, it is true that whilst cross-reactivity with human antibodies could help localise protein expression in *S. roscoffensis*, this does not mean that the antibodies are binding proteins that are true orthologues.

5.2.3.2 Anti-Nephrin and anti-Neph1 signal

No discernible signal was detected using commercial anti-Nephrin (anti-NPHS1, Novus Biologicals, AF4269) and anti-Neph1 antibodies (anti-Kirrel, Atlas Antibodies HPA030458), despite varying the concentration of primary antibodies (for protocol detail see section 2.9; Figure 5.11). It is clear from these images that endosymbiotic *T. convolutae* poses a problem in adult *S. roscoffensis* for non-specific fluorescent signal (Figures 5.11A and 5.11B). In juveniles, no signal was detected for anti-Neph1; anti-Nephrin signal appears to localise to a punctated pattern along the open 'edge' of the cup-shaped juvenile (Figure 5.11C).

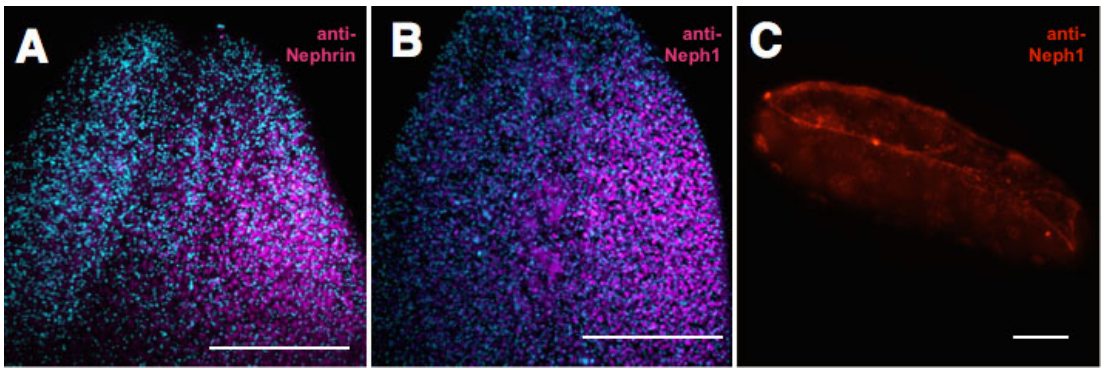


Figure 5.11. Immunohistochemistry using commercial antibodies against ultrafiltratory proteins on whole-mount *S. roscoffensis*. Anterior to the top of the panel in A and B, and to the left in C. Cyan or blue signal is DAPI. (A) anti-Nephrin in the adult; (B) anti-Neph1 in the adult. Magenta signal in both is autofluorescence from the endosymbiont *T. convolutae.* (C) anti-Nephrin in the juvenile, ventral surface at the top of the image. Signal appears to localise to the open cup of the ventral surface of the juvenile. Scale bars A and B: 100µm; scale bar in C: 50µm.

5.2.3.3 anti-Podocin signal

The commercial anti-Podocin antibody (anti-NPHS2, Atlas Antibodies HPA049486) appears to localise to two ladder-like bands of cells running in longitudinal stripes down the length of the adult animal (Figures 5.12A and 5.12C). This domain, in parenchymal cells flanking the midline of the gut, is comparable to that seen in *in situ* hybridisation experiments using sectioned adult *S. roscoffensis*. In juveniles, anti-Podocin signal is detected across the epidermis of the animal, with enhanced regions of signal along the lateral edges (Figure 5.12B). The very broad expression of anti-Podocin across the epidermis of the animal is likely to be non-specific signal, owing to the very conserved domains between Podocin/EB7/Mec2 and other stomatin-like proteins, as discussed in 5.2.5.

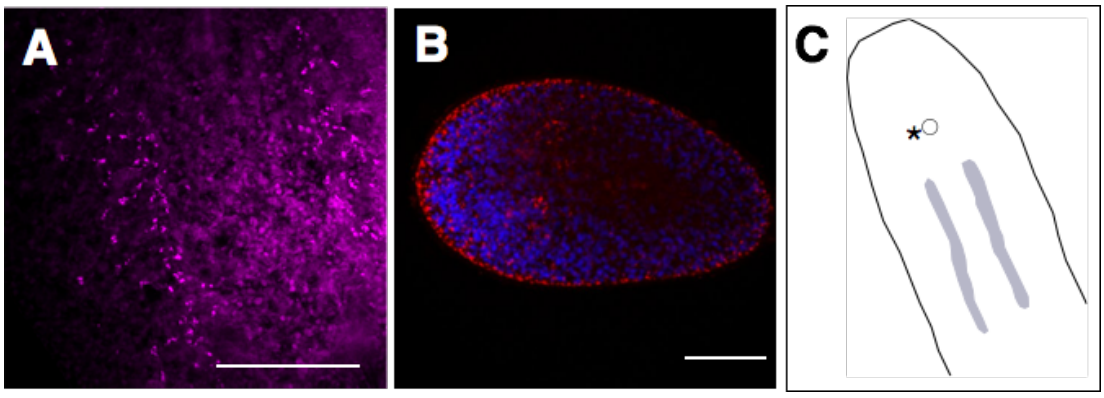


Figure 5.12. Immunohistochemistry using commercial antibodies against vertebrate Podocin in whole-mount *S. roscoffensis*. Anterior to the top in A; anterior to the left in B; cyan or blue signal counterstaining with DAPI. (A) anti-Podocin signal close to the midline of the animal presents as two longitudinal ladder-like bands of cells, on either side of the gut. (B) anti-Podocin in the juvenile: positive signal across the epidermis of the animal. (C) Context of anti-Podocin signal in the adult: anterior to top of schematic; statocyst represented by circle; location of anti-Podocin signal in grey. Scale bar in A: 100μm; scale bar in B: 50μm.

5.2.4 Immunohistochemistry: custom polyclonal antibody expression in *S. roscoffensis*

Given the inconclusive results using commercial antibody stainings in adult and juvenile *S. roscoffensis*, I hoped to gain a better understanding of the localisation of SrNeph1, SrNephrin and SrPodocin-like proteins by using specific polyclonal antibodies. Custom polyclonal antibodies were ordered from GenScript as described in 2.9.1.

5.2.4.1 Anti-SrNeph1 and anti-SrNephrin signal

Signal from anti-SrNeph1 and anti-SrNephrin was consistent in all experimental repeats. Signal from both antibodies appears to localise to the epidermal surface of the animal, in a net-like pattern (Figures 5.13A, 5.13B, 5.13F). From confocal Z-stacks, it is clear that this signal is localised to the epidermis of the animal; no specific signal for either antibody could be found in progressive Z-stacks through the dorso-ventral axis of the animal (Figure 5.13D and 5.13E). Double immunohistochemistry labelling found an overlapping signal for both anti-SrNeph1 and anti-SrNephrin in the epidermis (Figure 5.13C).

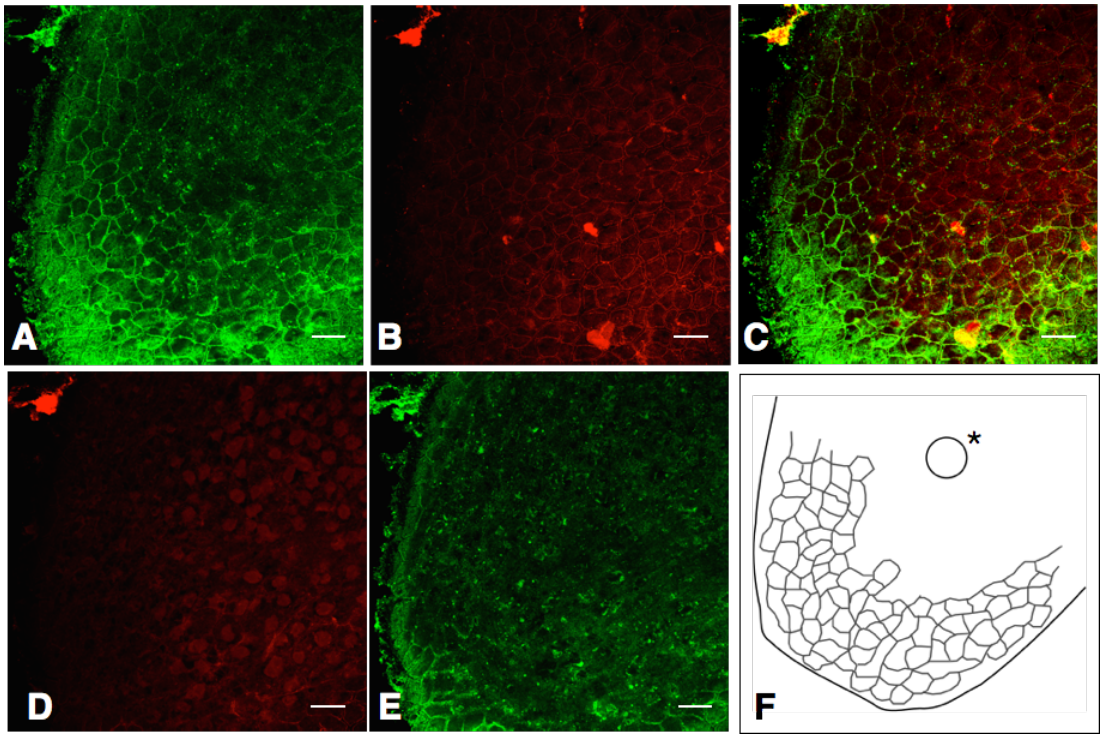


Figure 5.13. Signal from custom anti-SrNeph1 and anti-SrNephrin antibodies in the epidermal surface of whole-mount adult *S. roscoffensis*. Image taken from the anterior region of the animal (anterior tip in bottom left-hand corner of each panel). (A) anti-SrNephrin; (B) anti-SrNeph1; (C) Co-localisation of anti-SrNeph1 (green) and anti-SrNephrin (red). (D) and (E) Signal from custom anti-SrNeph1 and anti-SrNephrin antibodies in progressive Z-stacks through the dorso-ventral axis of whole-mount *S. roscoffensis*, image taken from the same anterior region of the same animal shown in figures A-C. (D) anti-SrNeph1 and (E) anti-SrNephrin. Scale bar in all aspects: 20µm. (F) context of anterior region shown in panels A-E, statocyst indicated by asterisk, signal from anti-SrNeph1 and anti-SrNephrin represented in grey.

In juvenile animals, anti-SrNephrin localises to the anterior-most section in the midline of the animal, with a broadening range of signal moving more posteriorly into the cup-shaped portion (Figure 5.14A). anti-SrNeph1 localises to the parenchymal region, running in longitudinal bands on either side of the animal, and in a lateral band across the midline of the animal, in the anterior-most third (Figure 5.14B). Double antibody stainings in juveniles show co-localisation across the epidermis, much like in adults, and with a possible co-localisation across the lateral band – although this is detected more strongly for anti-SrNeph1 (Figure 5.14C).

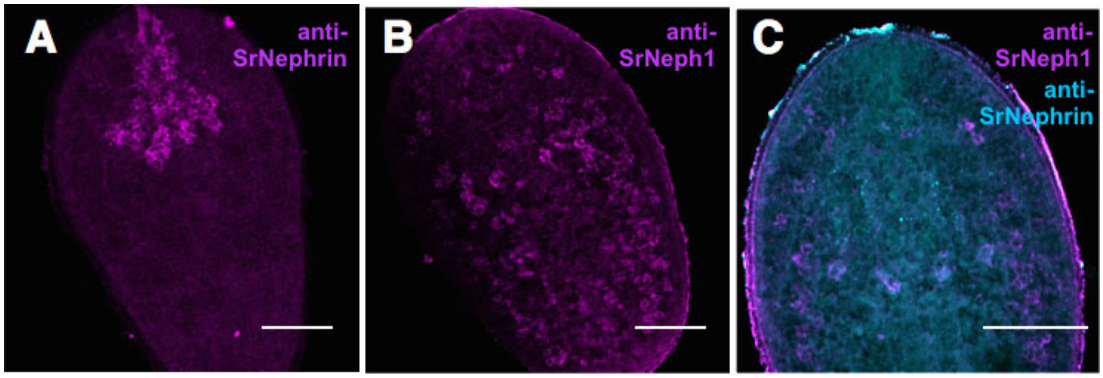


Figure 5.14. Signal from custom polyclonal antibodies anti-SrNeph1 and anti-SrNephrin in whole-mount juvenile *S. roscoffensis.* Anterior to the top in all aspects. (A) anti-SrNephrin. Signal at the anterior of the animal, with a broadening domain of signal moving posteriorly. (B) anti-SrNeph1, signal found in the parenchyma on either side of the gut, and in a lateral band across the midsection of the animal (C) co-localisation of anti-SrNeph1 (magenta) and anti-SrNephrin (cyan). Scale bar: 50μm.

5.2.4.2 Anti-SrPodocin-like signal

In both adults and juveniles, anti-SrPodocin-like gave a surprising signal in the early (juvenile) and fully-developed (adult) nervous system (Figure 5.15). Of all *S. roscoffensis* immunohistochemistry experiments, the signal from anti-SrPodocin-like was remarkably clear - and very consistent - across all experimental repeats. In juveniles, the antibody also localised to a punctated pattern around the edges of the animal, in the sagittocysts (Figure 5.15B).

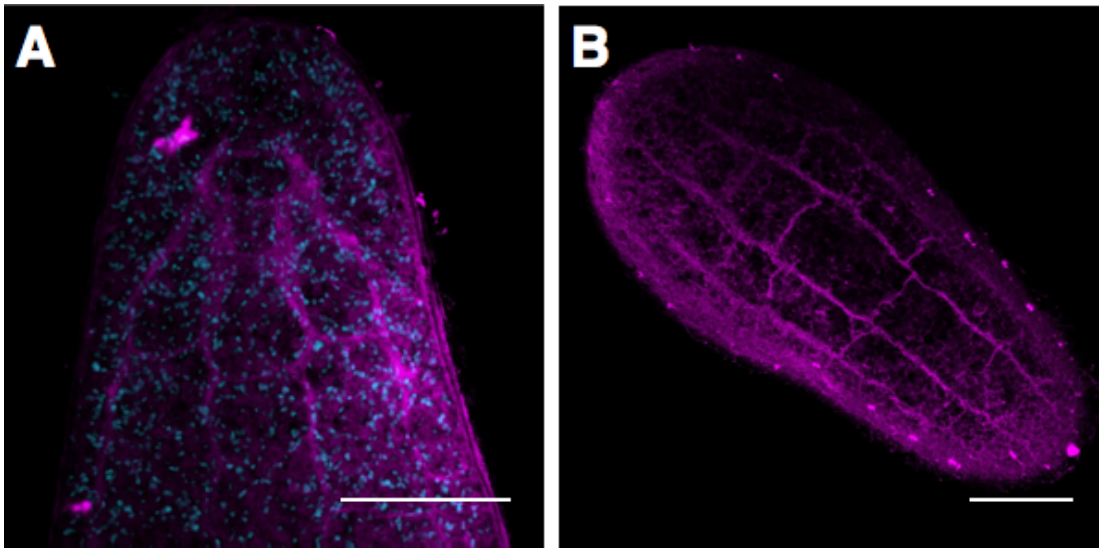


Figure 5.15. Signal from custom polyclonal antibody anti-SrPodocin-like in whole-mount *S. roscoffensis.* (A) Adult whole-mount; (B) Juvenile whole-mount. In A and B, signal localises to the nervous system. Anterior to the top in all aspects. Cyan signal in (A) is DAPI counter-staining. Scale bar in A: 100μm; B: 50μm.

5.2.5 Understanding expression patterns of ultrafiltratory genes in *S. roscoffensis*

It is clear that investigating the expression patterns of these three markers of ultrafiltration in the acoel *S. roscoffensis* was not straightforward. Both *in situ* hybridisation protocols used at both ages of fixed animal worked consistently to visualise the expression of *SrTroponin I*. This indicates that the difficulties encountered with *SrNeph1*, *SrNephrin* and *SrPodocin-like* probes were not owing to a protocol problem, but were specific to the RNA probes for the genes-of-interest. As there is no quantitative gene expression data for this species, the level these genes are transcribed at is not known in either the adult or the juvenile animal. It is therefore difficult to dismiss high levels of background staining in adult specimens as non-specific probe binding – particularly when antibody stainings indicate that *SrNeph1* and *SrNephrin* could be expressed broadly across the epidermis of the animal (Figure 5.13).

5.2.5.1 Problems with in situ hybridisation in S. roscoffensis and interpreting the validity of expression patterns

From expression patterns observed in juvenile specimens, *in situ* hybridisation experiments hint at the presence of ultrafiltration-related genes in broad expression domains in *S. roscoffensis*, including the parenchyma, gut, anterior and posterior regions. *In situ* hybridisation experiments in the nemertodermatid *Meara stichopi* and the acoel *I. pulchra* (Andrikou *et al*. preprint[158], Figure 5.16) have also found a diverse and inconsistent expression pattern for excretory system related genes: a result that is clearly evident for *S. roscoffensis*. Such broad expression could be indicative of a broad, non-filtratory specific function of these genes in Acoelomorpha members, but could also suggest non-specific probe binding as a result of the very conserved domains of CAM and stomatin-related proteins. *In situ* hybridisation in juvenile *S. roscoffensis* and the sectioned adult animal did consistently find expression of *SrNeph1* and *SrPodocin-like*, respectively, in parenchymal regions and close to the gut. When considered within the context of excretory systems, the presence of ultrafiltratory genes in parenchymal cells and the gut could be indicative of cells with a filtratory capacity contributing to excretion via the gut. Nonetheless, the problems encountered with *in situ* hybridisation experiments in *S. roscoffensis* makes this an unreliable result, and the data I present are not enough to inform our understanding of the roles of these genes in the Acoelomorpha.

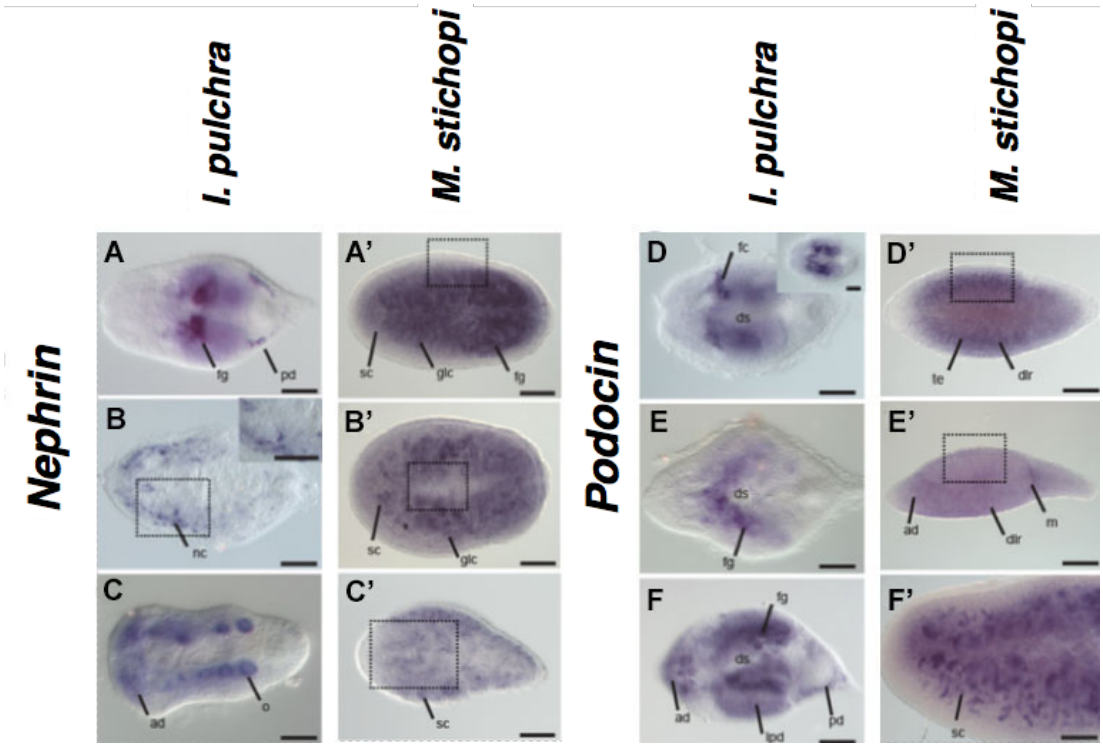


Figure 5.16. Expression of orthologues of ultrafiltratory-related genes in the acoel *Isodiametra pulchra* and nemertodermatid *M. stichopi.* Figure adapted from Andrikou *et al*. (preprint)[158] (A-C) *Nephrin* expression in *I. pulchra*; (A'-C') *Nephrin* expression in *M. stichopi;* (D-F) *Podocin-like* expression in *I. pulchra*; (D'-F') *Podocin-like* expression in *M. stichopi.* Scale bars A-C and A'-C': 50μm; D-F and D'-F': 100μm.

5.2.5.2 Signal from custom polyclonal antibodies

Using custom polyclonal antibodies in *S. roscoffensis* gave consistent and reproducible results. In adults, anti-SrNeph1 and anti-SrNephrin localise to a net-like structure across the epidermis of the animal (Figure 5.13). This signal is reminiscent of that previously described for a voltage-dependent anion-selective channel (VDAC) antibody used in *S. roscoffensis*, thought to show channels connecting monociliated sensory receptors[172]. In juveniles, different signal domains are observed in Z-stacks deeper into the dorso-ventral axis of the animal (Figure 5.14). anti-SrPodocin-like consistently localises to the nervous system (Figure 5.15).

As discussed in section 4.2.1, both Neph1 and Nephrin are evolutionarily conserved proteins found exclusively in the Bilateria. Outside of

filtratory diaphragm formation, both *Neph1* and *Nephrin* have been found to have a number of other developmental roles, including neuronal development and central nervous system patterning[135,173]. In *C. elegans*, orthologues of *Neph1* (*SYG1*) and *Nephrin* (*SYG2*) are expressed exclusively at the synapses[154]. Other members of the IgSF CAM family of proteins are also expressed in various neuron types in bilaterians[174]. Although Podocin is expressed exclusively at the podocyte slit diaphragm in vertebrates, the related stomatin proteins in mammals are also expressed in sensory neurons[175]. More widely in the Bilateria, *Mec2* and other stomatin members are commonly expressed in mechanosensory and touch receptors[176].

The expression of all three genes in neuronal cells of bilaterian organisms could explain the signal in *S. roscoffensis* using custom antibodies. However, the IgSF CAM family of proteins is very conserved at the level of their Ig-like extracellular protein domains (see section 4.2.1) and it is therefore possible that this signal is a result of cross-reactivity with other CAM proteins that are present in these cells. This would not rule out the potential true signal from SrNeph1 and SrNephrin in other non-neuronal cell types. Certainly for anti-SrNeph1 in juveniles, non-epidermal signal localises to cells found around the gut, in a pattern reminiscent of the expression of *SrNeph1* in *in situ* hybridisation. Similarly, Podocin/EB7/Mec2 proteins have exceptionally conserved domains with other stomatin family proteins, with consequences for both assigning orthology and for the likely cross-reactivity of anti-SrPodocin-like with other related proteins (see section 4.2.2). Whilst I consistently observed a strong nervous system signal for this antibody, the likelihood of detecting signal from related proteins for this antibody is high, and so the consistent nervous system signal can perhaps not be confidently assigned to SrPodocin-like.

5.3 General conclusions

S. roscoffensis has not been used extensively as a species for gene visualisation techniques. Where protocols have been used successively to date, these have predominantly been in juvenile specimens, and the problems I encountered in working with adult animals could be reflective of this trend. Of all adult *in situ* hybridisation results, the expression of *SrPodocin-like* in sectioned adult animals is likely to be the most reliable. This expression, in two longitudinal stripes along the midline of the animal, is reminiscent of the signal seen using the anti-Podocin commercial antibody. Whilst this expression could be from a paralogous gene, it is consistent between the two different visualisation approaches.

In situ hybridisation in juvenile *S. roscoffensis* gave more consistent results than those found using whole-mount adults. It appears that *SrNephrin*, *SrNeph1* and *SrPodocin-like* have broad domains of expression in juvenile animals, including the gut, parenchymal cells, and anterior and posterior regions. These genes are known to have a number of functions across the Bilateria (see section 4.2), and so broad expression domains would perhaps be likely. The common expression in the syncytial gut and parenchymal cells surrounding the gut could be indicative of an excretory function related to the gut and digestive cells. However the difficulties associated with implementing *in situ* hybridisation in this animal means that any expression patterns for the genes-of-interest should be interpreted with caution, and are not sufficient to draw conclusions regarding the ancestral roles of these genes or their function in acoels. Whilst the use of custom antibodies in immunohistochemistry gave consistent and reproducible results, the epidermal signal observed for anti-SrNeph1 and anti-SrNephrin could be attributed to cross-reactivity with related proteins. Signal from anti-SrNeph1 appears to correlate with that observed in *in situ* hybridisation experiments, and provides a degree of evidence for the expression of this gene in the gut and parenchymal cells in acoels. However, as discussed,

these results alone do not support the presence of specific nephrocyte-like cells with an ultrafiltratory capacity in *S. roscoffensis*.

It is evident that using *in situ* hybridisation and immunohistochemistry are not the best approaches to investigate the presence of cells with a putative ultrafiltratory function in *S. roscoffensis*. Instead, a functional approach could be preferable: RNAi has been used in the acoels *H. miamia* and *I. pulchra*[29,177]. In the flatworm *S. mediterranea*, RNAi of developmental transcription factors necessary for protonephridial patterning and maintenance resulted in visible bloating of the animal as a result of an inability to filter and excrete waste[34]. RNAi against *Neph1*, *Nephrin* and *Podocin* orthologues in acoels could therefore be used to investigate ultrafiltratory function of these genes with an easily scalable phenotype.

6 *Xenoturbella bocki* molecular protocols

6.1 Introduction

6.1.1 Visualising ultrafiltratory markers in *Xenoturbella bocki*

As described in section 4.1.3, identifying co-expression of *Neph1*, *Nephrin* and *Podocin-like* genes in the same cells is a likely indicator of ultrafiltratory function. Following the inconclusive *in situ* hybridisation and immunohistochemistry results for *S. roscoffensis*, I designed probes for the same three genes for use in *in situ* hybridisation in *Xenoturbella bocki* (subsequently referred to as *XbNeph1*, *XbNephrin* and *XbPodocin-like* respectively). The simple morphology of *Xenoturbella* means that it is commonly assumed that they lack any cells or structures involved in ultrafiltration and excretion. Identifying cells that co-express these ultrafiltratory genes in *Xenoturbella* would therefore be an important novel finding, with consequences for our understanding of the origins of ultrafiltration and nephridia, and of the degree of cellular specialisation in X*enoturbella*.

Nephridial systems in the Bilateria can broadly be divided into two classifications primarily determined by the presence (metanephridia) or absence (protonephridia) of the coelomic cavity for temporary filtrate storage[40]. Despite the structural variation in nephridial systems therein, the process of ultrafiltration is always mediated by the ECM (or basement membrane) prior to filtration by specialised epithelial cells in metanephridia or terminal cells in protonephridia. Consequently, in looking for cells in *Xenoturbella* that express genes associated with ultrafiltration, a location adjacent to the ECM might also be indicative of an ultrafiltratory role. In addition, a source of external or internally induced pressure is necessary to mediate ultrafiltration. In the vertebrate nephron filtration is driven by blood

pressure, and in protonephridia, internally-induced negative pressure is brought about by the beating of cilia within the terminal region, drawing fluid across the filtratory slits. In *D. melanogaster*, the source of filtratory pressure is more ambiguous, but appears to be reliant on the attachment of nephrocytes to peristaltic tissues such as the heart and gut[130].

6.1.2 Overview of *Xenoturbella bocki* morphology

First, I want to provide the morphological context necessary to interpret any positive results from *in situ* hybridisation or immunohistochemistry, included, as discussed, the location of the ECM.

All morphological and histological studies of *Xenoturbella bocki* provide evidence for a simple body plan. As originally described by Westblad (1949)[22], and supported by subsequent analyses, *Xenoturbella* comprise a small number of identifiable cell types, organised as layers through the body. These include a thick, ciliated epidermis - which includes the nervous system as a basiepithelial nerve net; an layer of extracellular matrix (ECM) underlying the epidermis; below this a muscle layer; a parenchymal layer; another thin ECM layer; and in the centre of the body, a gastrodermis surrounding the gut lumen[178] (Zakrzewski *et al.*, unpublished) (Figure 6.1). Beyond the basic organisation described above, coming from descriptive histological studies, little more is known regarding the possible variety of cellular subtypes or the diversity of gene expression (genetic fingerprints) that may allow us to characterise such cell or tissue type complexity in *Xenoturbella*.

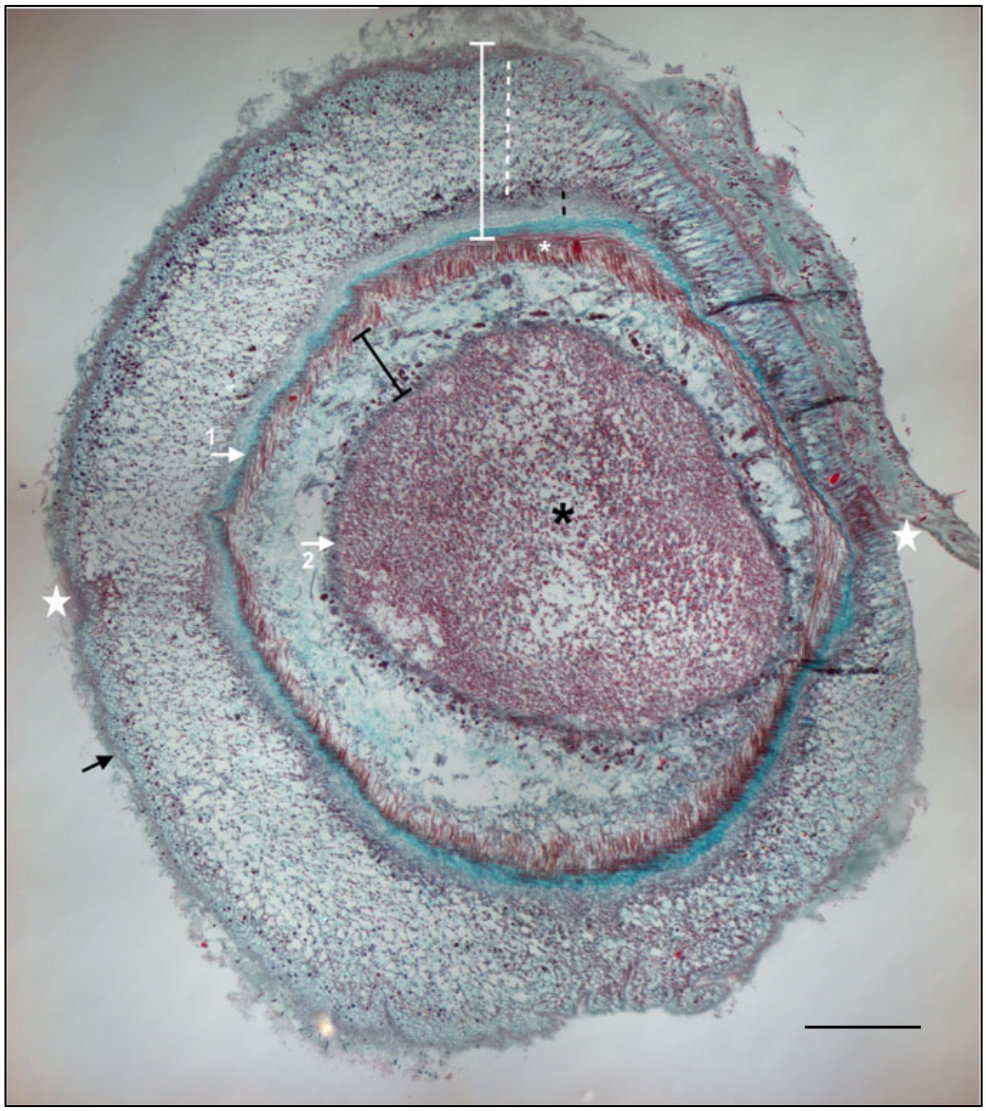


Figure 6.1. Masson's Trichrome staining of a horizontal cross section of *Xenoturbella bocki*. Section taken from the dorsal region of the animal (image Anne Zakrzewski, Telford lab). Outer ciliated epidermal layer shown by black arrow; thick epidermal layer encompassing the ciliated epithelia through to the turquoise ECM shown by white lines. The epidermis is a thick structure, comprising a number of different cell types. The outermost portion of the epidermis, shown by white dashed lines, is a monolayer, composed of cells up to 300μm in length and interspersed with gland, sensory, supportive, secretory and pigment cells. Underlying this portion of the epidermis is the nerve plexus (in grey), shown by black dashed lines. The cell bodies of neuron cells lie in the portion between the bulk of the nerve plexus and the epidermal monolayer. Underlying the nerve plexus, and in the most basal portion of the epidermis, is the extracellular matrix (ECM), stained bright turquoise and shown with the first white arrow. Unique to Xenoturbella and its close relatives (Xenoturbellida), a major part of the ECM is part of the subepidermal membrane complex (SMC): a thick layer of filaments surrounded by outer and inner basal laminae. The muscle layer itself is composed of three muscle types in distinct layers: the outer circular muscle,

the underlying longitudinal muscle layer, and a final transverse muscle layer which traverses the parenchymal space between the outer muscle layers and the gastrodermis. Muscle is stained red and annotated with a white asterisk. The parenchymal space itself (shown by black lines) is delimited on both sides by ECM – a thicker layer above the musculature, and a thinner inner layer shown by the second white arrow. This thinner layer of ECM is found adjacent to the gastrodermis – the cell layer found lining the large gut lumen, present as a sac in the middle of the animal (black asterisk). The two white stars opposite each other on the outside of the animal show the position of the ventral ring furrow. Scale bar: 200μm.

6.1.3 Objectives of chapter

In this chapter I will describe *in situ* hybridisation and immunohistochemistry experiments on sectioned adult *Xenoturbella* with the aim of visualising the expression of the three ultrafiltratory-related genes-of-interest described in section 4.1.3.

First, I describe the expression of a control probe for *Elav* (*embryonic lethal, abnormal vision*), a gene consistently associated with neuronal cell types across the Metazoa, from animals as diverse as vertebrates, *D. melanogaster* and the diploblast *Nematostella vectensis*[179-181], and for which a verified orthologue could be identified in *Xenoturbella* transcriptomic data (*XbElav* henceforth) by BLAST and subsequent phylogenetic analysis. A diffuse basiepithelial nerve net has been described from histological studies of *Xenoturbella*, and I therefore anticipated that this probe would localise to the nerve net, and act as reliable *in situ* hybridisation positive control gene during protocol establishment.

6.2 Results

6.2.1 Establishing the *in situ* hybridisation protocol with a test probe

Animals were sectioned along all three different body axes in order to optimise the interpretation of gene expression patterns across the whole animal (Figure 6.2):

1. Cross-sections, cutting vertically along the anteroposterior axis;
2. Sagittal sections, cutting vertically from left to right;
3. Horizontal sections, cutting along the dorsoventral axis.

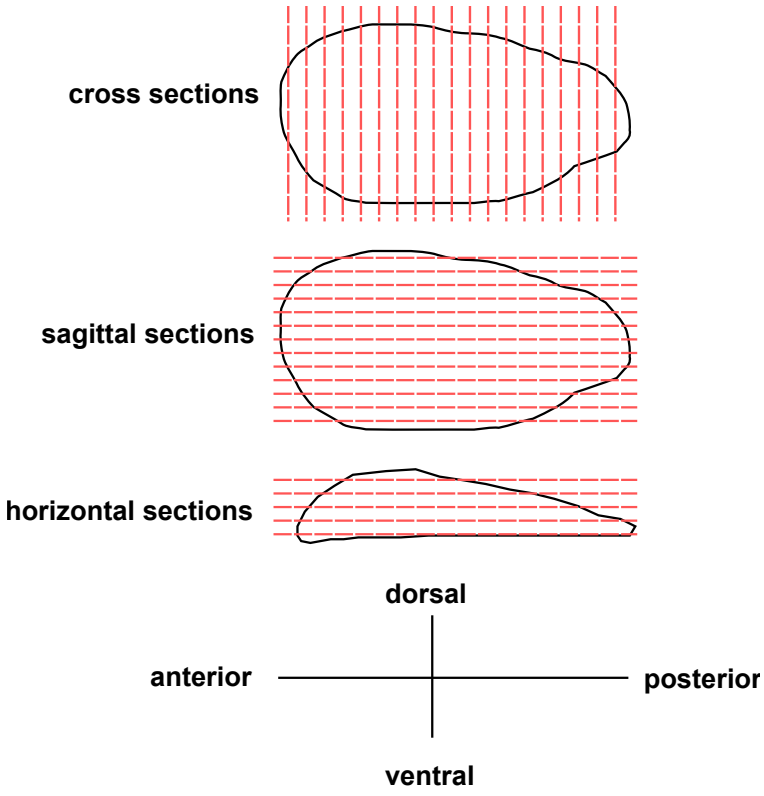


Figure 6.2. Sectioning orientations of whole-mount *Xenoturbella bocki.* Cross section and sagittal section viewed dorsally; horizontal section viewed laterally.

An initial round of *in situ* hybridisation on sagittal sections was carried out with a protocol modified from that originally used by Etchevers (2001)[182], using an RNA probe for *XbElav* (Figure 6.3). *XbElav* appears to show expression in the basiepithelial nerve net that is enhanced in intensity in the anterior region of the animal, similar to that observed in sagittal sections of the hemichordate *Saccoglossus kowalevskii*.

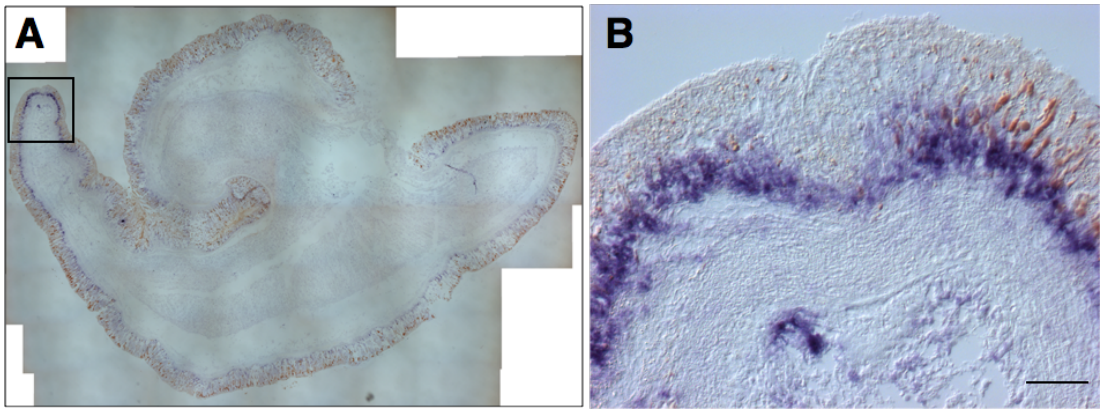


Figure 6.3. Test *in situ* hybridisation using a probe for *XbElav* on a sagitally orientated section of adult *Xenoturbella bocki*. (Experimental protocol and imaging in collaboration with Anne Zakrzewski, Telford lab). (A) Expression of *XbElav* appears to localise to the basiepithelial nerve net (corresponding to the cell layer indicated by the black dashed line in Figure 6.1). Anterior of animal to the left; large 'bulge' at the top of the image is a fixation artefact. Expression is strongest in the anterior-most region of the animal. (B) Detail of region indicated by black box. Expression of *XbElav* localises to the basiepithelial nerve net. Scale bar: 50μm.

6.2.2 Visualising expression of ultrafiltratory genes in *Xenoturbella bocki*

Owing to the lack of knowledge regarding distribution of cell types in *Xenoturbella*, *in situ* hybridisation experiments using probes for ultrafiltratory markers were first carried out on sections taken from all three sectioning orientations. Probes were synthesised as outlined in sections 2.6.3 - 2.6.5. For probe length and approximate location in sequence see Appendix 4.

6.2.2.1 XbNeph1 expression in the epidermis and posterior parenchyma

An initial set of *in situ* hybridisation experiments using a probe for *XbNeph1* showed a degree of expression across the thick epidermal layer, the muscle layer and the apical section of the gastrodermis (Figure 6.4A).

Despite the comparatively broad expression of *XbNeph1*, particularly strong expression was found in discrete cells at the border of the parenchyma with the gastrodermis (Figures 6.4B and 6.4C). These cells are clearly seen in the horizontally orientated section, where they appear to be settled on the ECM layer which overlays the basal portion of the gastrodermis (Figure 6.4C).

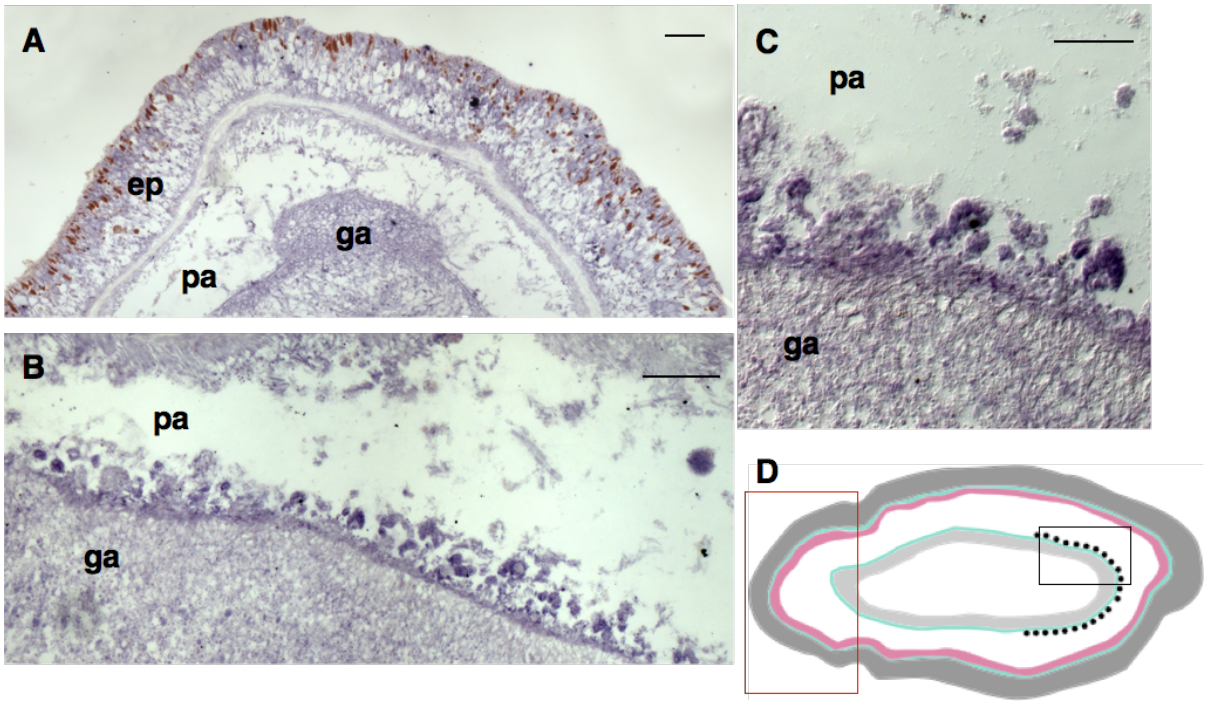


Figure 6.4. Expression of *XbNeph1* in differently orientated sections of *Xenoturbella bocki*. (A) Cross section showing the anterior-most region of the animal, expression of *XbNeph1* across the epidermal layer, in the muscle layer and in the gastrodermis; (B) Horizontal section, expression of *XbNeph1* strongest in the posteriorly located cells settling on the ECM; (C) Horizontal section showing detail of cells shown in (B); (D) schematic representation of *Xenoturbella*, with anterior to the left of the diagram. Dark grey represents the epidermal layer; two ECM layers shown in green; musculature shown in pink; gastrodermis in light grey; cells of interest lying adjacent to the gastrodermis shown at posterior of diagram. Red box to show region in panel A; black box to show regions in panels B and C. ep = epidermis, ga = gastrodermis, pa = parenchyma. Scale bars: A and B: 100µm; C: 50µm.

6.2.2.2 Specific XbNephrin expression in cells found in the posterior parenchymal region

Compared to *XbNeph1*, expression of *XbNephrin* is much less broad in the epidermal and gastrodermal cell layers. Some background staining can be seen in the basal-most portion of the gastrodermis and in the apical-most portion of the epidermis, but this expression is not seen in all sectioning orientations (Figures 6.5A and 6.5B).

Most interestingly, the strongest expression of *XbNephrin* is in specific parenchymal cells overlying the gastrodermal ECM, found at the posterior of the animal (Figure 6.5C). It is clear that *XbNephrin* has a degree less ubiquitous expression than *XbNeph1* in all orientations, and expression in these parenchymal cells is very localised.

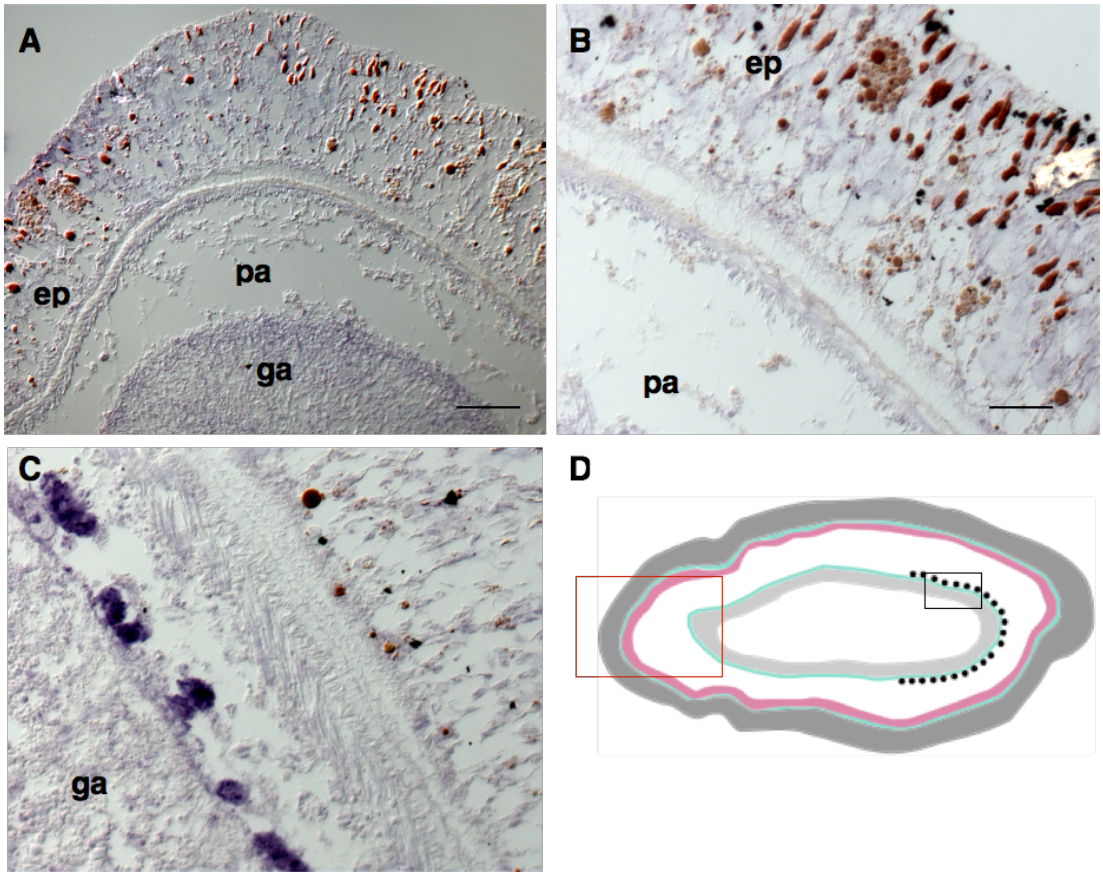


Figure 6.5. Expression of *XbNephrin* in differently orientated sections of *Xenoturbella bocki*. (A) Cross section showing the anterior-most region of the animal, faint expression of *XbNephrin* in the basal-most portion of the gastrodermis and the apical portion of the epidermis; (B) Horizontal section, faint expression of *XbNephrin* in the epidermis; (C) Horizontal section, strongest expression of *XbNephrin* in the posteriorly located cells settling on the ECM); (D) schematic representation of *Xenoturbella*, with anterior to the left of the diagram. Dark grey represents the epidermal layer; two ECM layers shown in green; musculature shown in pink; gastrodermis in light grey; cells of interest lying adjacent to the gastrodermis shown at posterior of diagram. Red box to show region in panel A and B; black box to show region in panel C. ep = epidermis, ga = gastrodermis, pa = parenchyma. Scale bars: A: 100µm; B and C: 50µm.

6.2.2.3 XbPodocin-like expression in the epidermis, gastrodermis and cells found in the parenchyma

Expression of *XbPodocin-like* across the epidermis and gastrodermis is weaker than that seen for *XbNeph1*. In cross section, some background staining can be seen in the basal region of the epidermal layer, lying apically to the nerve plexus but absent from the nerve plexus itself (Figure 6.6C). Expression can also be seen in the gastrodermal layer, surrounding the gut lumen, and in the parenchyma.

Compared to this background staining, the strongest expression of *XbPodocin-like* specifically localises to parenchymal cells settling on the ECM, overlying the gastrodermis (Figure 6.6A and 6.6B). Although there is a degree of *XbPodocin-like* expression in other parenchymal cells, it is evident that the highest level of expression is found in these cells, adjacent to the ECM, in the posterior region of the animal (Figures 6.6A and 6.6B). Importantly, these cells appear to be in the same location as the parenchymal cells that were found to positively express both *XbNeph1* and *XbNephrin*.

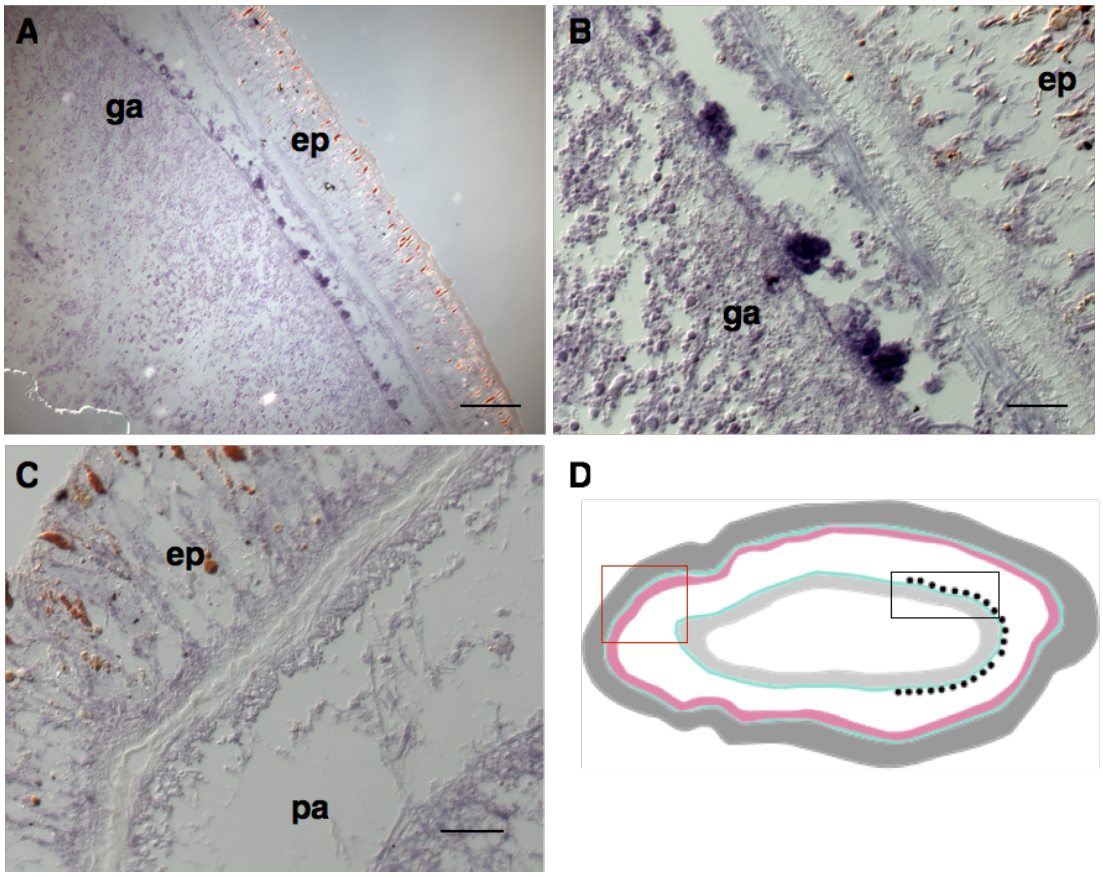


Figure 6.6. Expression of *XbPodocin-like* in differently orientated sections of *Xenoturbella bocki*. (A) Horizontal section showing expression of *XbPodocin-like* in cells found in the posterior region of the animal, settling on the gastrodermal ECM; (B) detail of cells shown in A (C) Cross section, showing faint *XbPodocin-like* expression in the epidermis and gastrodermal layer (D) schematic representation of *Xenoturbella*, with anterior to the left of the diagram. Dark grey represents the epidermal layer; two ECM layers shown in green; musculature shown in pink; gastrodermis in light grey; cells of interest lying adjacent to the gastrodermis shown at posterior of diagram. Black box to show regions in panel A and B; red box to show region in panel C. ep = epidermis, ga = gastrodermis, pa = parenchyma. Scale bars: A: 200µm; B and C: 50µm.

6.2.2.4 Expression of XbNeph1, XbNephrin and XbPodocin-like from initial in situ hybridisation experiments

Initial *in situ* hybridisation results using probes for *XbNeph1*, *XbNephrin* and *XbPodocin-like* found expression for all three genes in different domains within the epidermal layer. However, the expression of all three genes in cells in the parenchyma, settling on the thin ECM layer overlying the gut, is a particularly interesting finding that I wanted to investigate further. As these cells were most commonly identified using horizontally orientated sections of *Xenoturbella*, *in situ* hybridisation was repeated using the same three probes on horizontally orientated sections only.

6.2.3 Repeated in situ hybridisation on horizontal sections

As was evident from the first set of *in situ* hybridisation experiments, the expression of *XbNephrin*, *XbNeph1* and *XbPodocin-like* in the parenchyma is localised to specific cells that were consistently located in the posterior region of the animal. The horizontal sections used in this first set of experiments were taken from across the dorsoventral (DV) axis of *Xenoturbella*. Although the cells were observed for all three genes, this does not provide any specific information regarding the degree of distribution of these cells across the DV axis. To optimise the likelihood of identifying these cells in repeated *in situ* hybridisation experiments, a number of different successive horizontal sections were used for each probe, taken from across the DV axis.

6.2.3.1 XbNeph1, XbNephrin and XbPodocin-like are expressed in discrete cells overlying the gastrodermal ECM

Again, expression for all three genes was found to localise to distinct parenchymal cells, settling in the ECM overlying the gastrodermis, at the

posterior region of the animal (Figure 6.7). However, these cells are far less numerous in their distribution compared to the first set of *in situ* hybridisation experiments. Of the horizontal sections used for *in situ* hybridisation, a few cells with positive expression for each probe were found in sections taken from just one region of the DV axis, approximately ~3mm from the dorsal-most region of the animal, covering a total of ~140µm. Consequently, it appears that these cells-of-interest are possibly found as distinct clusters within the DV axis of this individual.

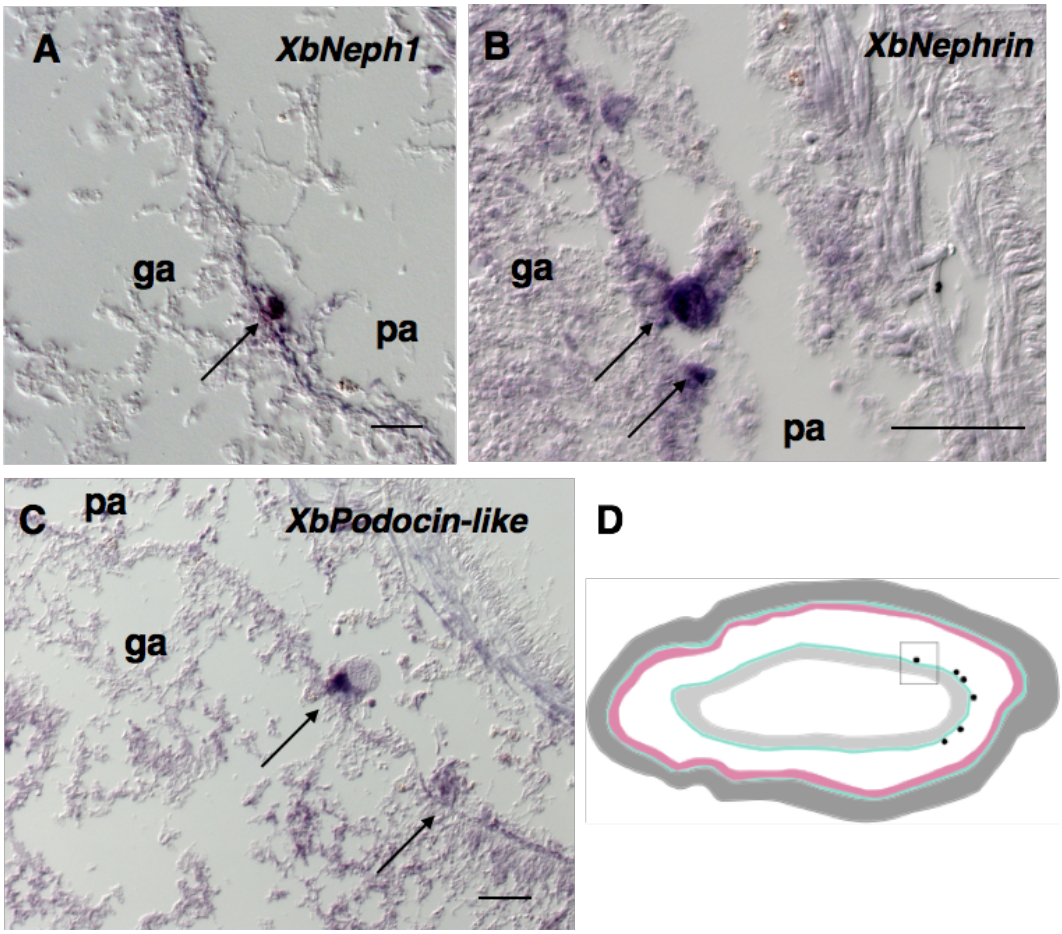


Figure 6.7. Expression of *XbNeph1*, *XbNephrin* and *XbPodocin-like* in posteriorly located cells overlaying the ECM on the basal side of the gastrodermis. Cells-of-interest indicated by black arrows. (A) *XbNeph1*, (B) *XbNephrin* (C) *XbPodocin-like.* Scale bars in all aspects 50µm. (D) Schematic representation of the location of these cells, with anterior to the left of the diagram. Dark grey represents the epidermal layer; two ECM layers shown in green; musculature shown in pink; gastrodermis in light grey; cells of interest lying adjacent to the gastrodermis shown at posterior of diagram. ga = gastrodermis; pa = parenchyma. Scale bars: 50µm in all panels.

6.2.3.2 Expression of XbNeph1, XbNephrin and XbPodocin-like in the epidermal layer

Repeated *in situ* hybridisation experiments focusing on horizontally orientated sections also recapitulated the expression pattern previously found in the epidermal layer. However, in this repeat, expression of all three genes in the epidermal layer is strongest at the location of the lateral sense furrow. *XbNeph1* expression is diffuse throughout the epidermal layer (Figure 6.8A), but strongest at the base of the epidermal monolayer, in cell bodies lying apically to the nerve plexus. Expression is particularly strong at the lateral sense furrow. No expression of *XbNeph1* is found in the basiepithelial nerve net itself. *XbPodocin-like* expression is seen strongly throughout the epidermal layer, but is most highly expressed at the base of the epidermal layer, apical to the nerve plexus (Figure 6.8C). *XbPodocin-like* expression is absent from the nerve plexus itself. *XbNephrin* expression is seen in the apical-most portion of the epidermis (Figure 6.8B), but is expressed very strongly in the cell bodies of the nerve plexus at the location of the lateral furrows.

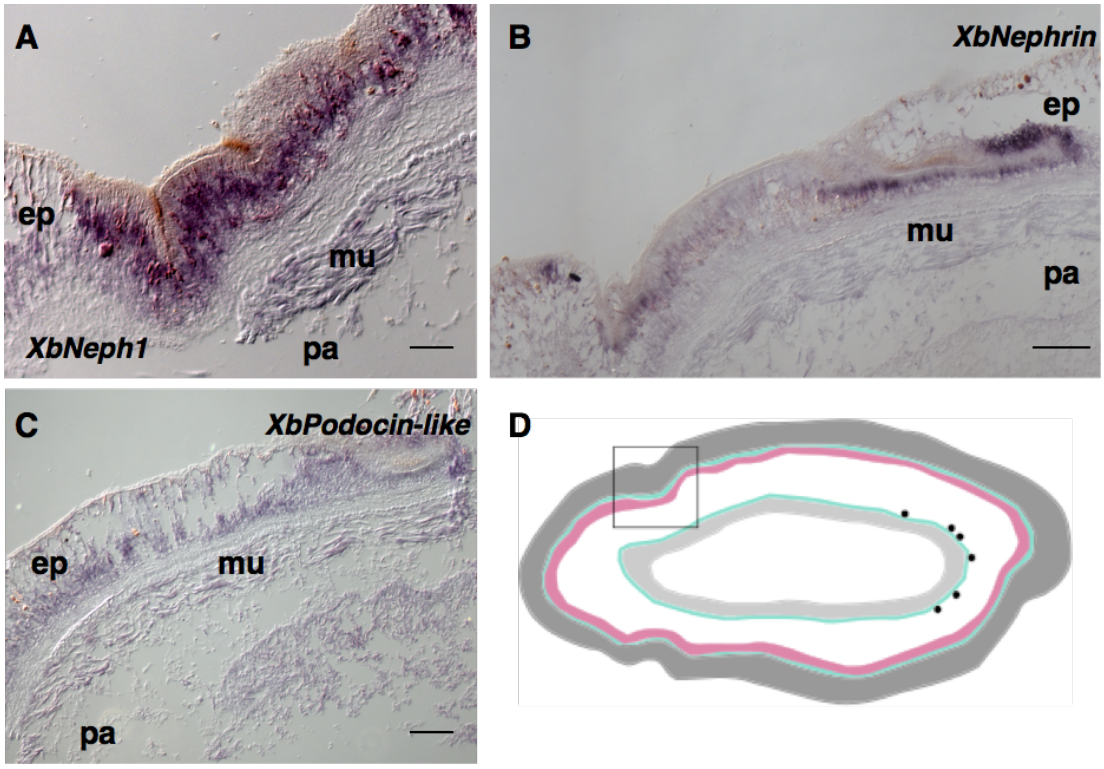


Figure 6.8. Expression of *XbNeph1*, *XbNephrin* and *XbPodocin-like* in lateral sense furrows at the anterior of *Xenoturbella*. (A) *XbNeph1*, (B) *XbNephrin* (C) *XbPodocin-like*. (D) Schematic representation of the location of the lateral sense furrow, with anterior to the left of the diagram. Dark grey represents the epidermal layer; two ECM layers shown in green; musculature shown in pink; gastrodermis in light grey; location of sense furrows shown in panels A, B and C indicated by black box. ga = gastrodermis; pa = parenchyma; mu = muscle. Scale bars: A and C: 50μm; B: 100μm.

6.2.4 Immunohistochemistry against ultrafiltratory proteins in *Xenoturbella*

Having used antibodies against Neph1, Nephrin and Podocin in *S. roscoffensis*, commercially made antibodies (anti-Neph1, anti-Nephrin and anti-Podocin, see section 5.2.3) and custom polyclonal antibodies raised against *S. roscoffensis* (anti-SrNeph1, anti-SrNephrin, anti-SrPodocin-like) were available for use in *Xenoturbella*.

Antibody stainings for the corresponding proteins-of-interest were carried out on sections already used for *in situ* hybridisation to look for co-localisation. Epitope sequences of commercial and the *Symsagittifera roscoffensis* custom polyclonal antibodies were compared to the inferred amino acid sequence for each of the three *Xenoturbella* genes, in order to choose the antibody that had closest similarity to the *Xenoturbella* sequence (*XbNephrin*: anti-NPHS1, Novus Biologicals, AF4269; *XbNeph1* and *XbPodocin-like*: Sr-custom antibody). However, as outlined in section 5.2.3, conservation between the respective antibodies and *Xenoturbella* sequences does not guarantee localisation of the polyclonal antibodies to a true orthologue in *Xenoturbella bocki*. Furthermore, as protein domains in Podocin and in Neph1 and Nephrin are well conserved with other Stomatin-like and CAM proteins, respectively, the possibility of cross-reactivity of these antibodies to other proteins is high (see sections 4.2.1 and 4.2.2). Consequently, the signal described for these antibodies in *Xenoturbella bocki* are not conclusive evidence for the presence of the protein-of-interest.

In situ hybridisation identified expression of *XbNeph1*, *XbNephrin* and *XbPodocin-like* in specific cells found settling on the ECM in the posterior parenchymal region of the animal. No signal for any antibody against the corresponding protein-of-interest was detected in cells in this location. However, all three antibodies (anti-SrNeph1, anti-Nephrin and anti-SrPodocin-like) appear to have domains of signal in the epidermis of *Xenoturbella.*

Signal from anti-SrNeph1 in *Xenoturbella* localises to cells in the upper-third of the epidermis, above the dense cluster of nuclei of nerve cell bodies indicated by DAPI staining (Figure 6.9A). Anti-Nephrin also localises to the epidermal layer, but the strongest signal is seen in the middle of the epidermal monolayer (Figure 6.9B). Anti-SrPodocin has two clear signal domains in the epidermis: a strong narrow band of signal in the ciliated cells at the apical-most portion of the epidermis, and diffuse signal throughout the basiepithelial nerve net (Figure 6.9C).

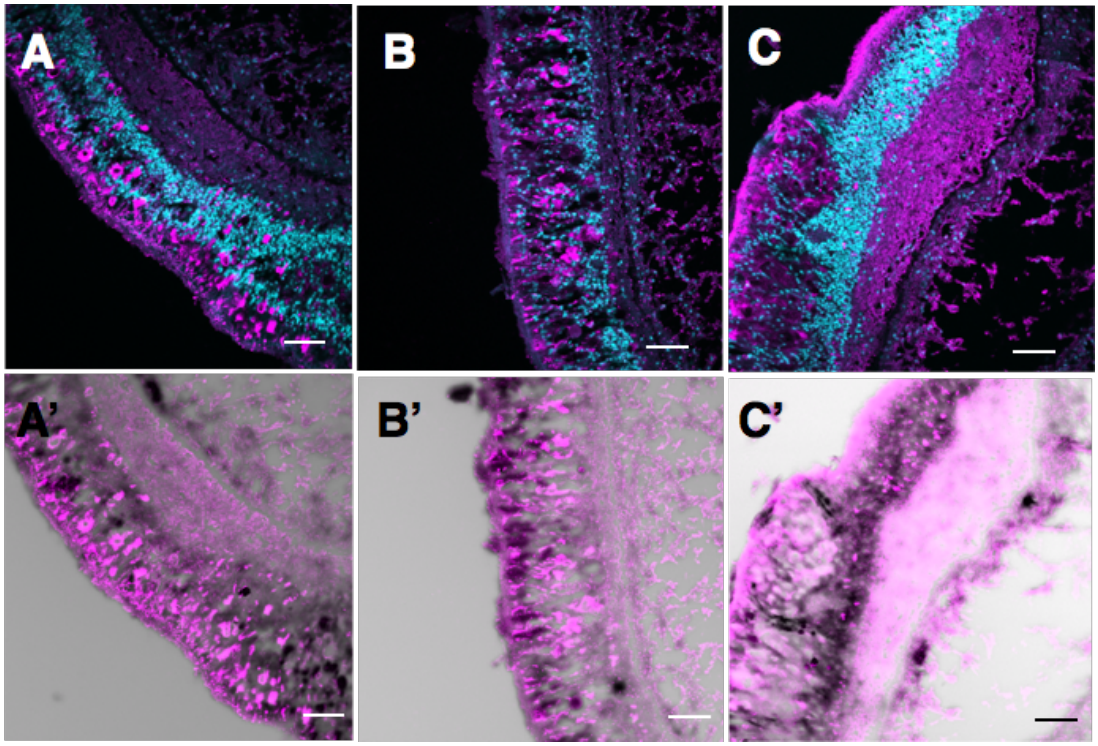


Figure 6.9. Antibody stainings using anti-SrNeph1, anti-Nephrin and anti-SrPodocin on horizontally orientated sections of adult *Xenoturbella.* Sections previously used for *in situ* hybridisation. In all aspects, epidermis is on the left of the panel, moving from left to right through the basiepidermal nerve net and into the gastrodermis and gut. A, B, C denote fluorescent signal from antibody stainings (magenta), counter-stained with DAPI (cyan); A', B', C' denote the same region overlaid with a DIC image to capture *in situ* hybridisation expression patterns. (A, A'): anti-SrNeph1 in *Xenoturbella*; (B, B'): commercial anti-Nephrin in *Xenoturbella*; (C, C'): anti-SrPodocin-like in *Xenoturbella.* Scale bars in all aspects: 100μm.

6.3 Discussion

6.3.1 Reliability of the *Xenoturbella in situ* hybridisation protocol

The *in situ* hybridisation protocol established on sections of adult *Xenoturbella* appears to give consistent results. The expression of the positive-control *XbElav* within the basiepithelial nerve net, and enhanced in the anterior-most region, is in line with that expected for *Elav* as a metazoan pan-neuronal marker[179-181]. Furthermore, *in situ* hybridisation using RNA probes against a number of putative cell-type specific genes, identified in the *Xenoturbella* single cell sequencing protocol and described in Chapter 7, show specific expressions in expected cell or tissue types. Consequently, I am confident that the expression patterns found for ultrafiltratory-related genes in *Xenoturbella* are showing expression for specific RNA probes. Conversely, the non-specificity of the antibodies used in immunohistochemistry mean that the signal seen from anti-Nephrin, anti-SrNeph1 and anti-Podocin in *Xenoturbella* is less reliable.

6.3.2 Ultrafiltratory gene expression in posteriorly located cells overlying the gastrodermis

Probes for all three genes-of-interest show expression in cells in the posterior lateral side of the animal, in close proximity to the ECM lining the gastrodermis. These are individual cells that appear to be settled on the ECM (Figure 6.7). *XbNeph1*, *XbNephrin* and *XbPodocin-like* all appear to be expressed in these cells despite a degree of background staining in the gastrodermal layer, which is particularly evident for *XbNeph1* and *XbPodocin-like*. As described in section 4.1.3, the common cellular expression of *Neph1*, *Nephrin* and *Podocin* is widely accepted to be a marker of ultrafiltratory function. In the vertebrate kidney, molecular interactions between adjacent podocyte cells (Nephrin-Nephrin homodimers and Neph1-Nephrin heterodimers) are necessary for the formation of the ultrafiltratory

diaphragm[135]. Ultrafiltration in *D. melanogaster* is carried out in discrete nephrocyte cells, with infoldings within the nephrocyte cell itself leading to the physical formation of the ultrafiltratory nephrocyte diaphragm[130]. Filtered waste material is subsequently sequestered within the nephrocyte cells, prior to metabolism or excretion via the Malpighian tubules.

The expression of these three ultrafiltratory genes in discrete cells lying adjacent to the gastrodermis in *Xenoturbella bocki* could be indicative of an ultrafiltratory function for these cells in an organism that is widely-regarded to lack ultrafiltratory or excretory specialisation. In all bilaterian excretory systems, the ECM acts as a primary filtratory barrier: the presence of these cells adjacent to the ECM also provides a degree of evidence for a possible ultrafiltratory role. Furthermore, the ECM layer underlying these cells is found basal to the epithelial gastrodermis - the cellular layer lining the large central gut lumen of *Xenoturbella*. In the absence of a coelomic cavity, the presence of ultrafiltratory cells directly overlying the gut is not unlikely: the position of these cells lying in close proximity to the ECM of the gastrodermis would imply that if such cells did function in ultrafiltration, then filtration would occur moving out of the gut and through the ECM, before being sequestered within these cells.

Whilst *in situ* hybridisation experiments appear to show an intriguing common expression of *XbNeph1*, *XbNephrin* and *XbPodocin-like* in these posteriorly positioned cells, this result needs further verification as evidence for the presence of nephrocyte-like cells or cells with an ultrafiltratory capacity in *Xenoturbella bocki*. Sections taken from just two different adult *Xenoturbella* were used for *in situ* hybridisation, but it appears that these cells are variable in number across the DV axis of the animal and between individuals. Given their sparse distribution, *in situ* hybridisation on further sections is needed to confirm their presence. Nonetheless, the lack of knowledge regarding specific cell types in *Xenoturbella bocki* means that the identification of these cells is an intriguing finding. Alternative approaches for investigation into their gene expression and distribution – for example, using TEM in conjunction with immunogold labelling, which would necessitate

species-specific antibodies - may also help to shed light on their function. Furthermore, as has been described, the nephrocyte cells in *D. melanogaster* are found attached to peristaltic tissue, necessary for the induction of ultrafiltration. *Xenoturbella* possesses a simple blind gut, but muscle fibres have been found to traverse the parenchymal region and attach to the gut epithelium (Figure 6.1). To what degree these muscle fibres could mediate pressure for filtration is unclear, but elucidating a source of ultrafiltratory pressure would be an important consideration for their function.

6.3.3 Epidermal and neuronal expression of putative ultrafiltratory genes in *Xenoturbella bocki*

The epidermis in *Xenoturbella* is a thick cell layer comprising a number of different cell types, with histological studies indicating that it is a more complex structure than it appears at first glance (Zakrzewski *et al.,* unpublished). All three probes of interest localise to the epidermal layer (Figure 6.8), with strongest epidermal expression in the location of the anteriorly positioned lateral furrows. Whilst these furrows have previously been described as sensory related furrows (Zakrzewski *et al.,* unpublished), their sensory function remains to be confirmed. Nonetheless, DAPI staining at the location of the lateral furrows shows a very dense cluster of nuclei lying apically to the basiepithelial nerve net (Figures 6.9A and 6.9C), in what is thought to be the cell bodies of the nerve plexus (Figure 6.1). The presence of more nerve plexus cell bodies, pertaining to more axons at this location, could therefore provide a degree of support for a sensory or neuronal-related function at these anterior furrows. As described in section 5.2.5, Neph1, Nephrin and Podocin/Stomatin-like genes are known to have roles in neuronal/nerve function and patterning[135,154,173,175,176], and so the expression of these genes in this putative neuron-rich location could be indicative of a conserved neuronal cell function.

Polyclonal antibodies also appear to give an epidermal signal (Figure 6.9). For *XbNeph1* and *XbPodocin-like*, the epitope of the respective

Symsagittifera roscoffensis custom polyclonal antibody was more similar to the inferred *Xenoturbella* protein sequence; for *XbNephrin*, the commercial antibody sequence (anti-NPHS1, Novus Biologicals, AF4269) was more similar. Nonetheless, as outlined in section 6.2.2, using polyclonal antibodies raised against different species does not guarantee the identification of an orthologous protein.

When overlaid with the corresponding DIC image, there appears to be a degree of co-localisation of anti-SrNeph1 signal with *XbNeph1* expression in the cell body layer overlying the nerve plexus (Figures 6.9A and 6.9A'). However, signal from anti-SrNeph1 is also strongly detected in the apical-most portion of the epidermis, which was not found for the *XbNeph1* probe. This fluorescent signal is clearly localised to cells that have been hypothesised to function as gland or secretory cells, and which stain red/orange in *in situ* hybridisation images as a result of the *in situ* hybridisation protocol (for example Figure 6.9A). This signal is not seen as strongly for anti-Nephrin or anti-SrPodocin-like antibodies, and could therefore be signal from a CAM-protein member.

The strongest anti-Nephrin signal is found in the mid-section of the epidermal cell layer (Figure 6.9B), in the same region corresponding to *XbNephrin in situ* hybridisation expression, most clearly visible in Figure 6.8B. Some anti-Nephrin signal is also found in the apical-most portion of the epidermal layer, but in fewer cells than that found for anti-SrNeph1. This does not correlate with that observed in *in situ* hybridisation: in the DIC image (Figure 6.9B'), dark staining in the apical region of the epidermal layer instead corresponds to the red/orange colouration of putative gland cells.

Signal from anti-SrPodocin-like in *Xenoturbella bocki* is most different from that found for the *XbPodocin-like* RNA probe. Anti-SrPodocin localises most strongly to the basiepithelial nerve net itself (Figure 6.9C), with reduced signal in the neural cell body layer. Strong signal from anti-SrPodocin-like is also evident in the cilia of the epidermis. *Xenoturbella bocki* is known to move by gliding on motile cilia, but the possibility presence of primary

sensory cilia, also on the epidermis, remains to be investigated. *Mec2*, the *C. elegans* orthologue of Podocin/EB7 functions in mechanosensory cells[176], and it is possible that a *Podocin-like* gene could similarly function in mechanosensation in *Xenoturbella*, but this cannot be concluded from the SrPodocin-like signal seen here.

6.4 General conclusions

Using *XbElav* as a positive-control neuronal probe shows clearly that *in situ* hybridisation has been established successfully for the first time in *Xenoturbella bocki*. In addition, immunohistochemistry has been implemented on slides previously used in *in situ* hybridisation. As so little is known regarding gene expression domains in *Xenoturbella*, these gene visualisation approaches are valuable tools for contributing to our understanding of their morphology and organisation.

Investigating the expression domains of *XbNeph1*, *XbNephrin* and *XbPodocin-like* found intriguing results, which although not conclusive, hint at newly discovered cell types in *Xenoturbella*. RNA probes for *XbNeph1*, *XbNephrin* and *XbPodocin-like* appear to show common expression in discrete cells overlying the gastrodermis at the posterior of the animal. As outlined in section 6.3.2, this is far from conclusive evidence for an ultrafiltratory cell specialisation, but it does identify unique cell types that will prove informative for further investigation. Furthermore, their location, settling on the ECM overlying the gastrodermis, is consistent with an ultrafiltratory function.

In addition, the epidermal and neuronal expression of RNA probes for these genes and signal from corresponding polyclonal antibodies are comparative to the anti-SrNeph1, anti-SrNephrin and anti-SrPodocin-like signal seen in *Symsagittifera roscoffensis* (see section 5.2.4). Although cross-reactivity of antibodies between species does not guarantee identification of true orthologues, these findings do suggest that signal from

antibodies against the proteins-of-interest localises to the epidermis and putative neuronal cells in the Xenacoelomorpha: a finding that is supported by the function of all three genes in neuronal cell types in other bilaterian organisms[135,154,173].

Whilst these results could suggest a conserved neural function of *Neph1*, *Nephrin* and *Podocin-like* genes in xenacoelomorphs, the identification of discrete cells in the parenchyma that appear to express *XbNeph1*, *XbNephrin*, and *XbPodocin-like* is a more interesting finding. Owing to protocol establishment for single *in situ* hybridisation in *Xenoturbella*, the co-expression of these genes has yet to be confirmed with a double staining protocol, but this would certainly be informative for localising expression in the same cell. It would also be beneficial to carry out future *in situ* hybridisation experiments on sections taken from more individuals of different sizes, to better understand how the number and distribution of cells varies between animals.

Alternative microscopy approaches such as TEM could also help to elucidate the structure of these cells, and contribute to our interpretation of them as potential ultrafiltratory structures: a finding that would have implications for our classification of the Nephrozoa and the origin of filtratory and excretory systems.

7 Single cell sequencing

7.1 Introduction

7.1.1 Gene expression in *Xenoturbella bocki*

7.1.1.1 Using in situ hybridisation to identify putative cell types

As detailed in Chapter 6, I have been successful in establishing and implementing *in situ* hybridisation and immunohistochemistry protocols on sectioned *Xenoturbella.* This means that, for the first time, we can investigate the expression patterns of genes that are associated with a specific function or cell type, or genes that are known to have stereotypical expression across the Bilateria[183]. In the context of searching for nephrocyte-like cells in *Xenoturbella*, I have used *in situ* hybridisation to identify the expression of three ultrafiltratory-related genes (*XbNeph1*, *XbNephrin* and *XbPodocin-like*) in cells positioned on the gastrodermal ECM at the posterior of the animal. More broadly, *in situ* hybridisation could be used to visualise the expression of genes in *Xenoturbella* that can help us to understand more about cell type complexity in the animal, and where different cell types might be distributed across the body.

Whilst *in situ* hybridisation is a valuable approach to have established, the huge number of potential genes for investigation and the very little we currently know about *Xenoturbella* body organisation means that selecting the best candidate marker genes for *in situ* hybridisation can be problematic. As samples are a limiting factor for *Xenoturbella*, selecting the most appropriate genes for investigation is essential in order to optimise use of the sections that are available. *In situ* hybridisation experiments are time-consuming and in particular they are difficult to scale-up to investigate more than two genes simultaneously in the same section.

7.1.1.2 Characterising cell types by their transcriptional profiles (single cell sequencing)

Another recently developed approach to achieve the same goal – and much more besides - is to characterise individual cells by their transcribed genes, to identify subsets of cells with closely matching transcriptional states (cell types, for example, nephrocyte, muscle, and nerve cells) and subsequently to use limited *in situ* hybridisation to locate these cells in the animal. Because the cell types have been deeply characterised in terms of transcriptional state, a single *in situ* hybridisation can provide context for many of the genes expressed in the identified cells.

Single-cell RNA sequencing (scRNA-seq) is a high-throughput, quantitative method to assess cell state specific transcriptomes. In single cell sequencing, RNA-Seq libraries are prepared from dissociated cells taken from either a specific tissue[184-186] or from across the whole organism[187]. Computational clustering approaches are then applied to the single cell libraries to group cells together into meta-clusters based on a common transcriptional profile. By identifying highly expressed genes within the meta-clusters, putative identities can be assigned to these cells. As I will show, this information can then be used to identify optimal markers to enable *in situ* hybridisation to locate these cells within the animal.

As described, I have used *in situ* hybridisation to identify putative cells that appear to co-express genes that are commonly associated with ultrafiltration. Using single cell sequencing in *Xenoturbella* could be used to identify cells that are enriched for these genes – and in this way to identify molecular markers of ultrafiltration – as validation of my current *in situ* hybridisation results, and as a way of deeply characterising these cells or identifying other specific and highly expressed markers. More broadly, whole organism single cell sequencing in *Xenoturbella* presents a high-throughput technique for understanding gene co-expression and identifying distinct cell states in the animal.

In the following section I outline how transcriptional profiles can be used to identify putative cell or tissue types. I then describe how single cell sequencing could be used in *Xenoturbella* to complement current nephrocyte-related *in situ* hybridisation results, and how clustering single cell libraries can contribute to what is currently known regarding cell type complexity and gene expression in *Xenoturbella*.

7.1.2 Cell types and the specification of tissues

7.1.2.1 Cell types across the Metazoa

Multicellular organisms (Metazoa) are characterised by the presence of more than one cell type, defined as any cell with unique physiological or structural characteristics[188]. Animals comprise different cells specialised for discrete functions: non-bilaterians including the cnidarians, ctenophores and poriferans have historically been considered to have relatively few cell types, and cell type number has commonly been used as a marker for organism complexity. Different cell types themselves have also been used as a tool to infer evolutionary relatedness. For example, the presence of ultrafiltratory and excretory related cell types in the so-called Nephrozoa (and their assumed absence in the Xenacoelomorpha) has been used to provide a degree of evidence for the position of the Xenacoelomorpha outside of the main bilaterian Nephrozoa grouping (see section 1.3).

7.1.2.2 Physiological and molecular identifiers of cell types

Traditionally, identifying shared cell types between different lineages was based on cell morphology[189]. Cellular characteristics were compared using microscopy-based approaches: a strategy that was helpful in identifying homologous cell types between closely related species, but more ambiguous when applied to distinct animal phyla[188]. A more reliable marker of cellular homology comes from comparing cellular characteristics at the molecular

level. Cell types can be defined – be that structurally or physiologically - by a unique combination of expressed differentiation genes and regulatory transcription factors. Transcriptional profiles can thus be compared across species to identify putatively closely related cell types in different lineages[188,190,191].

In nephridial systems, as we have seen, a number of genes have been identified that are commonly expressed at the site of ultrafiltration in diverse bilaterian taxa (see section 4.1.3). Consequently, identifying cells based on their transcriptional profile that co-express all or a combination of these genes (*Neph1*, *Nephrin*, *Podocin*, *CD2AP*, *ZO-1*) is a likely indicator of ultrafiltratory specialisation. More broadly, cells that comprise specific tissue types are characterised by the common expression of known genes – for example, actin and myosin in muscle cells. By comparing the transcriptional profiles of different cells within the same tissue type, we can even begin to identify heterogeneity or diversity within a broader cell-type grouping. For example, identifying striated muscle cells based on the specific expression of troponin members, or smooth muscle cells based on calponin expression. Identifying cells that are characterised by these conserved molecular markers across diverse organisms can inform our understanding of how the diversity of cell types evolved, and allow us to infer which cell types might be evolutionarily ancient, or present in eumetazoan or bilaterian ancestors.

7.1.3 RNA-Seq and single cell sequencing (scRNA-Seq)

7.1.3.1 Using scRNA-Seq to identify cell and tissue types

The identification of distinct cell types has been significantly aided by recent developments in single cell sequencing technology. The most widely used single cell assay approach is RNA-Seq, which measures gene expression by reverse transcribing RNA into cDNA and sequencing these molecules: the number of reads originating from each gene gives a measure of its level of expression in a given cell or tissue. RNA-Seq was originally developed for investigating multi-cellular populations[191] (individual cells have

very little mRNA). However, assigning cell type identity based on bulk cell analysis is confounded by a number of problems. Cells are dynamic entities, and so even cells taken from the same population are likely to have varied transcriptional states. Unless a truly homogenous and synchronised cell population is sequenced, bulk measurements averaged across the population will hinder the reconstruction of a cell state specific transcriptional profile[191]. In order to derive a meaningful molecular fingerprint of a cell type, it is necessary to sequence the RNA complement of individual cells. Implementing RNA-Seq and generating cDNA libraries from picograms of RNA has required significant technical refinements, but has been successfully applied to tissues derived from model organisms. This has provided insight into, for example, defined cell types in the mouse brain[185]; in the developing midbrain of human and mouse[184]; and in mouse retinal neurons[186], amongst others.

7.1.3.2 Whole organism single cell sequencing in Xenoturbella

The first applications of whole organism single cell sequencing have been carried out in the cnidarian *Nematostella vectensis* (Sebé-Pedrós *et al.*, in review)[192], using a single cell protocol and analysis pipeline called MARS-Seq[88], and in the well characterised model organism *C. elegans*[187]. For non-model organisms such as *Xenoturbella*, single cell sequencing presents a powerful tool for investigating cell type complexity in animals which have not been extensively studied, and for which relatively little is known about their body organisation. Applying whole organism single cell RNA-Seq to *Xenoturbella* presents a high-throughput approach to look for cells that co-express the ultrafiltratory genes-of-interest, but also to group cells together into putative cell or tissue types, based on common gene expression profiles.

7.1.4 Objectives of chapter

The objective of this chapter was to apply the MARS-Seq single cell sequencing protocol[88] to whole *Xenoturbella* adults. Based on the single cell

libraries, I hoped to identify cell types based on common transcriptional profiles. With this in mind, I aimed to group cells together into meta-clusters based on common gene expression patterns. By identifying highly expressed genes within these clusters, I hoped to assign putative identities to these cell states. Based on this, I aimed to use *in situ* hybridisation experiments to validate the single cell sequence approach and assign a spatial profile for each of the cell populations identified. In addition, I hoped to identify cells in the single cell libraries that co-expressed the ultrafiltratory related genes-of-interest as a complementary approach to the *in situ* hybridisation results outlined in Chapter 6.

7.2 Results

Single cell sequencing of *Xenoturbella* was carried out during collaborative visits with Heather Marlow and lab members at the Pasteur Institute, Paris. For methodological detail regarding cell sorting and library preparation see sections 2.10.1 and 2.10.2, and for computational analysis of the single cell libraries see section 2.10.5. Computational analysis of the data was predominantly carried out in collaboration with Yann Loe Mie (Marlow lab, Pasteur Institute).

7.2.1 Success of the single cell pipeline in *Xenoturbella bocki*

Using the MARS-Seq sequencing and analysis pipeline originally established to study red blood cell transcriptomes[88], single cell sequencing was successfully implemented in *Xenoturbella*. A total of 6080 single cell libraries were prepared, made from RNA extracted from 3040 cells from each of two adult animals (8x 384 well plates per animal, with four empty control wells per plate).

Unlike previous tissue-specific applications of single cell sequencing, whole organism approaches result in much greater transcriptional variation

between cell types. The sequencing library data from *Xenoturbella* showed a broad variability in cell size and total RNA-content between individual cells (Figure 7.1A). Furthermore, there was a very wide variation in the number of Unique Molecular Identifiers (UMIs) contributed by individual genes. By including a UMI as a random DNA sequence label, each molecule in the final cDNA library is made unique. After sequencing, each UMI might be observed multiple times, but the original number of DNA molecules can be inferred by grouping identical UMIs together and counting this as a single instance of the read. This gives a reliable method for identifying unique transcripts. When the amount of starting RNA is small (as with single cell sequencing), UMIs can also be used to infer the relative starting size of different samples. Whilst some UMIs in the single cell libraries mapped only once to a given gene model, the most highly expressed gene, *actin gamma 1-like* (*ACG1-like*) contributed >8000 UMIs to the dataset (Figure 7.1B).

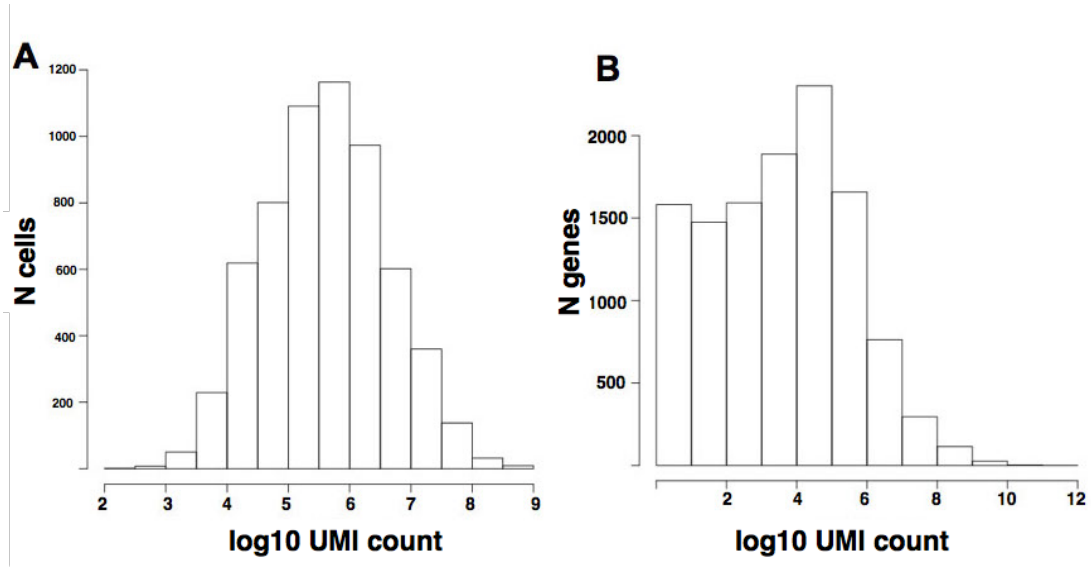


Figure 7.1. Wide distribution of number of unique transcripts (UMIs) per cell and per gene in *Xenoturbella.* (A) Distribution of log10 UMIs per cell. (B) Distribution of log10 UMIs per gene.

7.2.2 Searching the single cell libraries for cells that co-express the ultrafiltratory genes-of-interest

Primarily, I investigated the single cell data for cells that might commonly express the ultrafiltratory genes-of-interest. All 6080 single-cell libraries were searched for expression of the ultrafiltratory genes described in section 4.1.3 (*XbNeph1*, *XbNephrin*, *XbPodocin-like*, *XbCD2AP* and *XbZO-1*). From this, just nine cells were identified as expressing *XbNeph1*, and 45 cells as expressing *XbNephrin* (Figure 7.2). The expression of *XbCD2AP*, *XbZO-1* and *XbPodocin-like* was higher, in 140, 48 and 356 cells, respectively. No cells were found to co-express all five ultrafiltratory markers, but cells were identified as co-expressing different combinations of the ultrafiltratory genes (for example, five cells co-expressing *XbPodocin-like* and *XbNephrin*, and two cells co-expressing *XbPodocin-like*, *XbNephrin* and *XbCD2AP*, Figure 7.2).

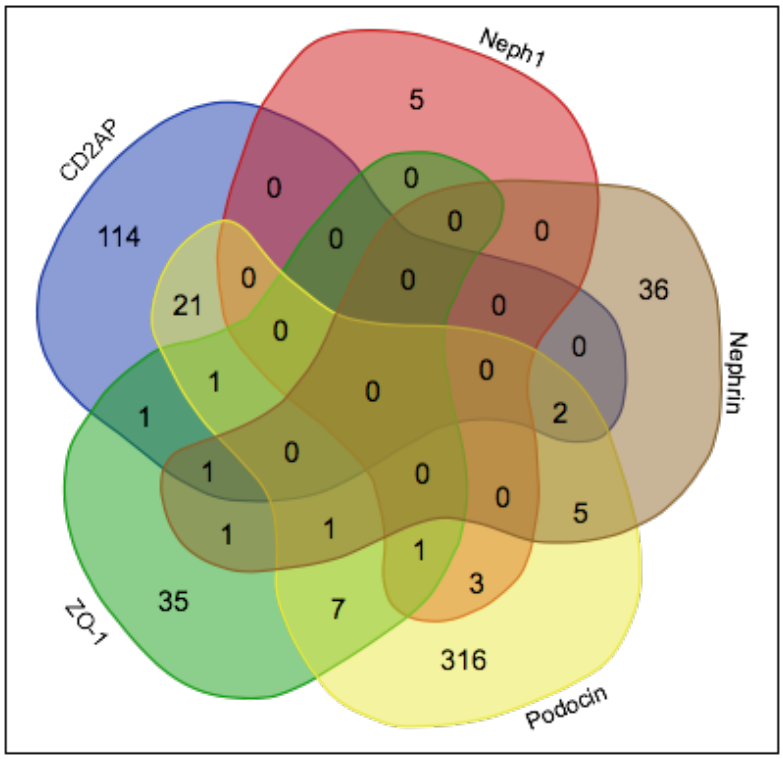


Figure 7.2: Venn diagram showing limited overlapping cell-specific expression of ultrafiltratory related genes. Diagram shows expression of *XbNeph1*, *XbNephrin*, *XbCD2AP*, *XbPodocin-like* and *XbZO-1* from all 6080 *Xenoturbella* single cell libraries.

7.2.3 Clustering the single cell libraries using Seurat

My second objective of single cell sequencing in *Xenoturbella* was to identify as many putative cell types as possible (to go well beyond the search for nephrocytes) in the single cell libraries. This can be done without *a priori* specification of which cell types to look for. In order to identify groups of cells that are characterised by a common transcriptional signature – that is, enriched expression of the same genes - the Seurat[89] pipeline for clustering single cell data was used on the *Xenoturbella* single cell libraries. Seurat is an unsupervised machine-learning approach to cell clustering: single cell libraries are grouped by common transcriptional signatures, and not by *a priori* specification of the assumed diversity of the data. For detail of the Seurat clustering protocol see section 2.10.5.

The Seurat cell clustering output is visualised as a tSNE plot (t-Distributed Stochastic Neighbor Embedding). tSNE is a machine learning algorithm for interpreting complex, high-dimensional data in two or three dimensions, the results of which can be represented on a scatter plot. For the *Xenoturbella* single cell libraries, Seurat tSNE analysis was constructed based on 15 dimensions of variance in the data. In the tSNE plot, cells of the same colour (where one circle represents one cell) are grouped together based on a common transcriptional profile. As can be seen on the tSNE plot, Seurat grouped cells into 16 transcriptionally distinct populations, but with a degree of overlap within the four clusters found at the centre of the plot (Figure 7.3). This indicates a degree of common gene expression found between cells in these clusters.

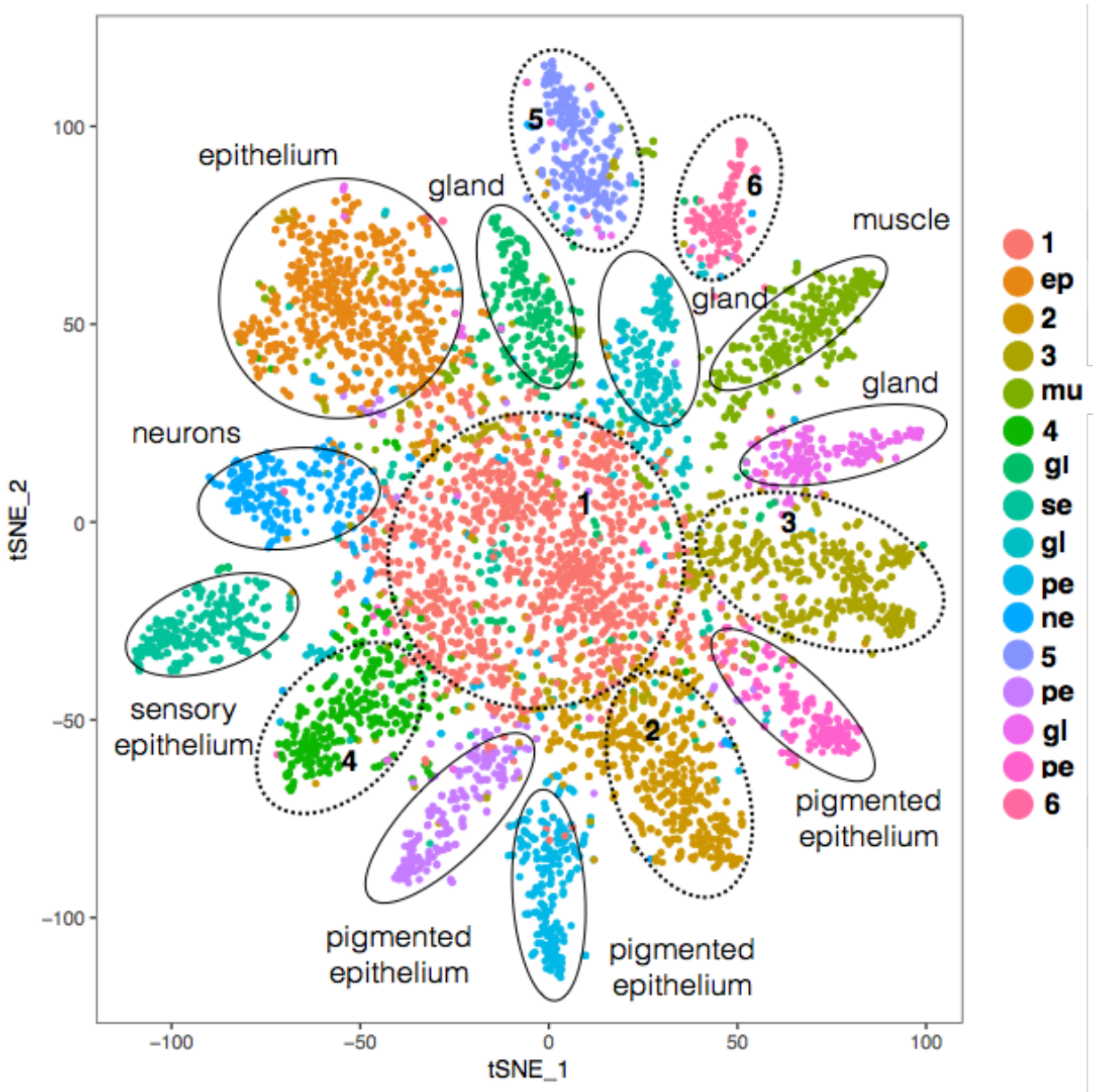


Figure 7.3. tSNE plot generated using Seurat, showing spatial distribution of cell meta-clusters. Meta-cluster identity assigned to epithelial (ep); muscle (mu); gland (gl); sensory epithelium (se); pigmented epithelium (pe); and neural (ne) cell populations, based on transcriptional profiles described in section 7.2.4. Unassigned meta-clusters indicated by dashed lines and numbers 1-6.

7.2.4 Identification of meta-clusters in the single cell data

Using the Seurat clustering pipeline, 16 different meta-clusters of cells were identified, grouped together based on commonly expressed genes. In order to assign putative identities to each of the meta-clusters, blastx queries were used to identify highly expressed genes within each meta-cluster. Where the highest similarity on NCBI was to an uncharacterised or hypothetical protein, genes were assigned as uncharacterised; if no sequence similarity could be found, genes were assigned as not annotated (NA).

For ten of the meta-clusters, highly expressed genes were identified that have well conserved, cell-type specific expression in diverse metazoan taxa. Based on the expression of these annotated genes, six transcriptional profiles were assigned, associated with epithelial cells, digestive/secretory glands, muscle, neurons, sensory/neural epithelia and pigmented epithelia. For example, the muscle meta-cluster was assigned based on enrichment for myosin and troponin members; the neural meta-cluster identified by upregulation of Neurotrophin and Synaptotagmin 1-like; and putative digestive gland cells identified based on the expression of Chymotrypsin-like genes and various proteases, amongst others.

As is clear from the tSNE plot, some of these putative cell-types are represented by more than one meta-cluster (Figure 7.3). Muscle, epithelial, sensory epithelial and neuronal cell meta-clusters are represented only once, whilst putative gland and pigmented epithelial cells are represented by three meta-clusters (Figure 7.3). This indicates a degree of transcriptional heterogeneity across these cell-type classifications. Confident cell-type identity could not be assigned to the four diffuse clusters of cells represented in the centre of the tSNE plot (Clusters 1-4, Figure 7.3), or to clusters labelled as 5 and 6 in the tSNE plot (Figure 7.3).

7.2.5 Diversity within meta-clusters

Genes that were enriched in specific meta-clusters were visualised on the tSNE plot. For some meta-clusters, cell-specific expression of genes revealed diversity within a broader cell-type classification. For example, in the neural meta-cluster, mapping the expression of different genes onto the cells in the tSNE plot revealed a potential diversity in neural types. Some genes, including Synaptotagmin 1-like and Secretagogin have a diffuse and broad expression across the neural meta-cluster, with a degree of upregulation of Synaptotagmin 1-like also seen in the sensory epithelial cluster (Figure 7.4). Others, including a *Xenoturbella*-specific gene, are found only in certain cells within the broader neural meta-cluster (red 'NA' gene in Figure 7.4). This could be indicative of a degree of neural cell type diversity. Similarly, expression of digestive-related enzymes, including Chymotrypsin-like and a protease gene is seen in cells within one of the gland cell meta-clusters. A *Xenoturbella*-specific gene (blue 'NA' gene in Figure 7.4) is expressed in two different gland cell meta-clusters, which suggests possible heterogeneous functions across the gland cell groupings.

Conversely, the muscle meta-cluster appears to be largely homogenous for the upregulation of muscle-related genes, such as troponin and myosin (Figure 7.5). Genes that have upregulated expression in the epithelial cell meta-cluster commonly also have broader expression across other meta-cluster cell-types (for example, Innexin and Olfactomedin, Figure 7.5).

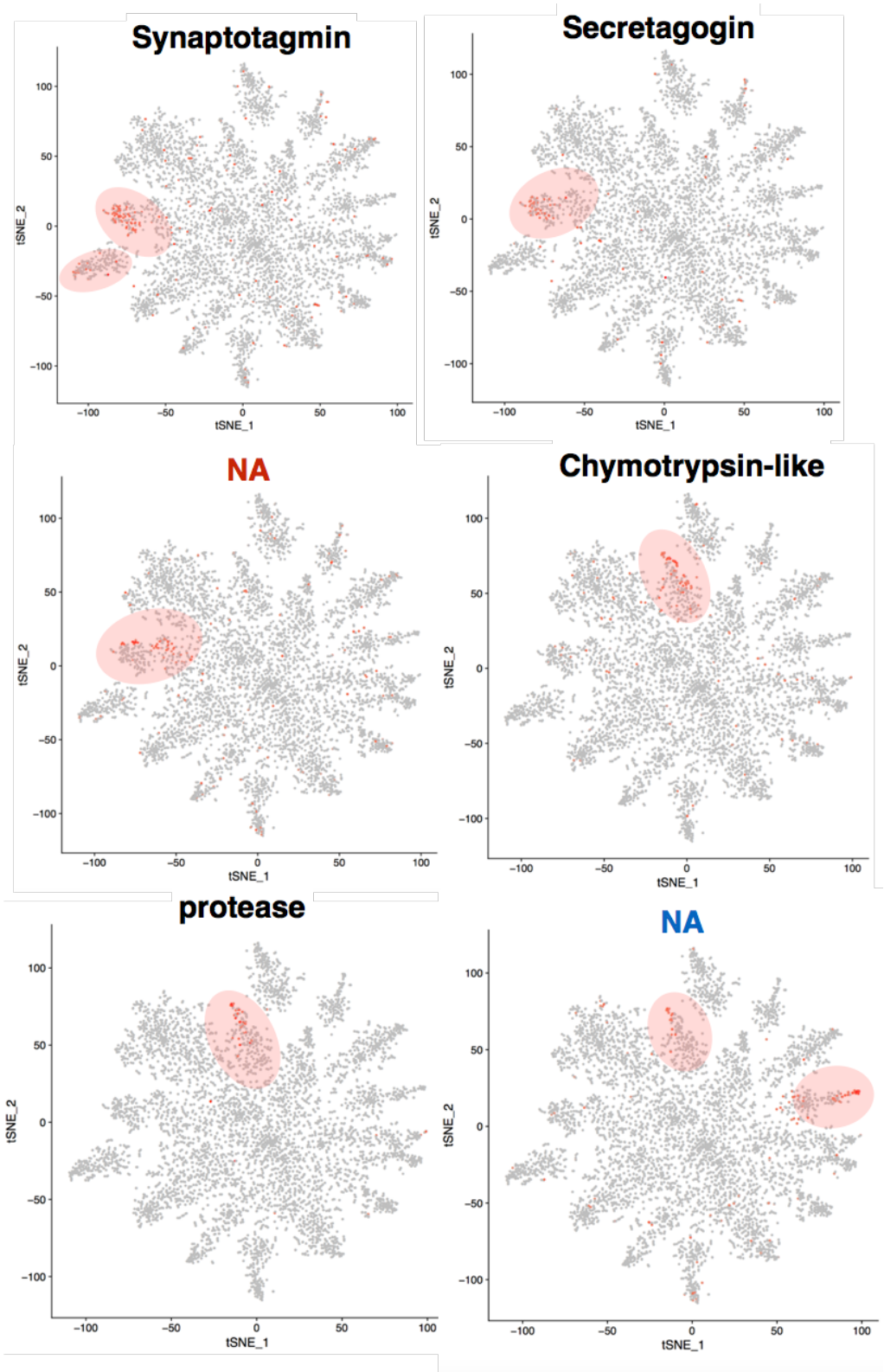


Figure 7.4. Mapping of selected genes expressed in the neural and gland cell meta-clusters onto the Seurat tSNE plot. Synaptotagmin 1-like, Secretagogin and an uncharacterised *Xenoturbella*-specific gene (in red) (putative neural identity). Chymotrypsin-like, a protease member and a different uncharacterised *Xenoturbella*-specific gene (in blue) (putative gland cells).

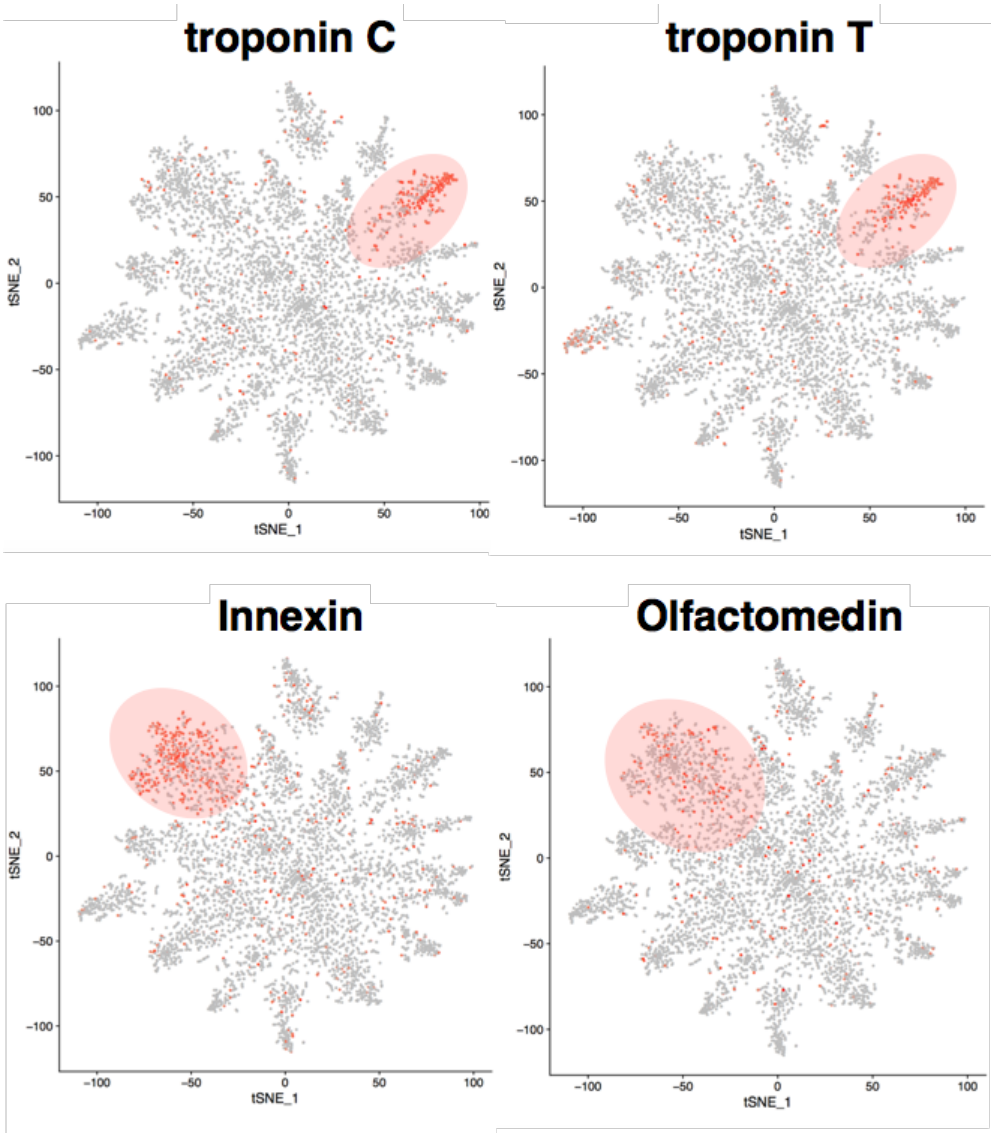


Figure 7.5. Mapping of selected genes expressed in the muscle and epithelial cell meta-clusters onto the Seurat tSNE plot. Troponin (muscle meta-cluster); Innexin and Olfactomedin (putative epithelial cells).

For all meta-clusters, unannotated, *Xenoturbella*-specific genes (NA genes) were amongst those with the most specific meta-cluster expression.

7.2.6 *In situ* hybridisation validation of putative meta-clusters

To validate the classifications assigned to six of the putative cell meta-cluster types, *in situ* hybridisation using probes for genes with specific meta-cluster enrichment were carried out on horizontally sectioned adult *Xenoturbella*.

Primarily, *in situ* hybridisation was carried out using probes for muscle genes (troponin members). Prior morphological and histological analysis shows that muscle cells in *Xenoturbella* are found in a defined layer lying basally to the epidermal nerve net and underlying ECM. Identifying the expression of these known muscle markers strongly and specifically in this defined cell layer therefore acted us a promising positive control for single cell data validation (Figure 7.6).

Subsequent *in situ* hybridisation experiments focused on other known meta-cluster specific markers. Chymotrypsin-like, identified as a gland-specific gene, is found expressed in cells of the gastrodermis, providing evidence for a digestive-specific function (Figures 7.7A and 7.7B, see Appendix 8 for orthology assignment). Tyrosinase-like 1 is expressed uniquely in cells found in the epidermal layer, which, given the prominent pigment spots seen across the body of *Xenoturbella*, is a likely domain of expression for putative pigmented cells (Figure 7.7C). For the epithelial meta-cluster, a probe for Innexin showed strong expression throughout the basiepithelial nerve net (Figure 7.7D).

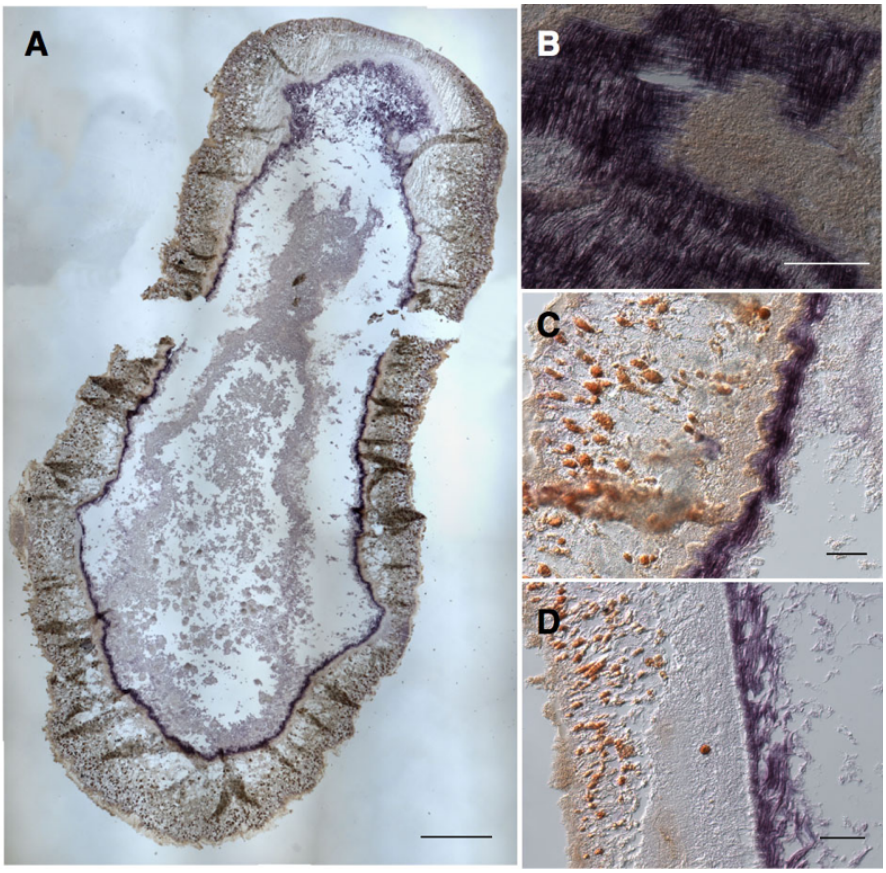


Figure 7.6. *In situ* hybridisation validation of Troponin T and Troponin C expression in the muscle cells of *Xenoturbella bocki* (A) Overview of Troponin T expression in a horizontal section of adult *Xenoturbella*. Anterior at the top of the panel; (B) Detail of expression of Troponin T in the circular and longitudinal muscle; (C) Detail of Troponin T expression in the muscle layer, underlying the epidermis and epithelial nerve net; (D) Detail of Troponin C expression. Scale bars: A: 200µm; B,C,D: 50µm.

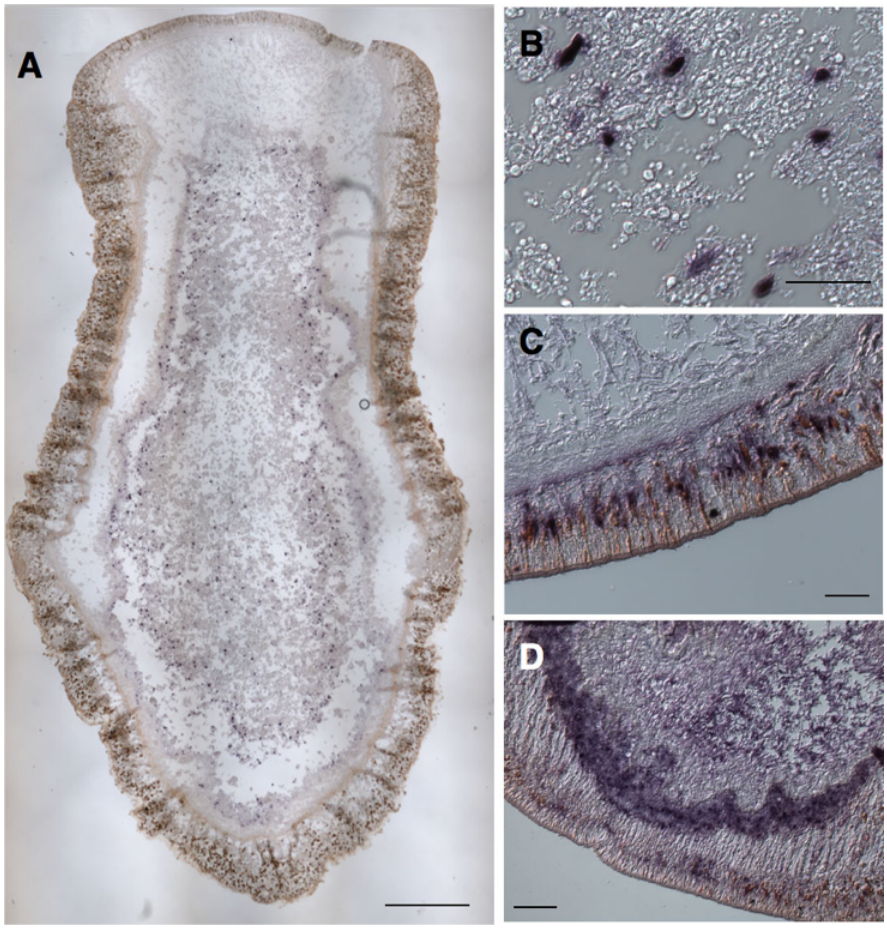


Figure 7.7. *In situ* hybridisation validation of annotated genes in *Xenoturbella bocki* single cell libraries. (A) Overview of Chymotrypsin-like expression in a horizontal section of adult *Xenoturbella*. Expression is localised to the gastrodermis of the gut. Anterior at the top of the image, section taken from the dorsal side of the animal, on the edge of the gastrodermis; gut lumen in the centre of the animal. (B) Detail of Chymotrypsin-like expression in the gastrodermis. (C) Tyrosinase-like 1 expression in the pigmented cells of the epidermis (D) Innexin expression in the basiepithelial nerve net. Scale bars: A: 200µm; B, C: 50µm; D: 100µm.

Most interestingly, *in situ* hybridisation using a probe against a *Xenoturbella*-specific gene hypothesised to be expressed within a subset of the neuronal cells (expression on the tSNE plot shown as the red NA gene in Figure 7.4) showed expression in discrete cell bodies found uniquely in the anterior epidermal region of the animal (Figure 7.8). The localised expression of this gene in these anteriorly located cells provides spatial context for an interesting neural-related gene, and demonstrates the benefits of using single cell sequencing results to guide probe selection for *in situ* hybridisation.

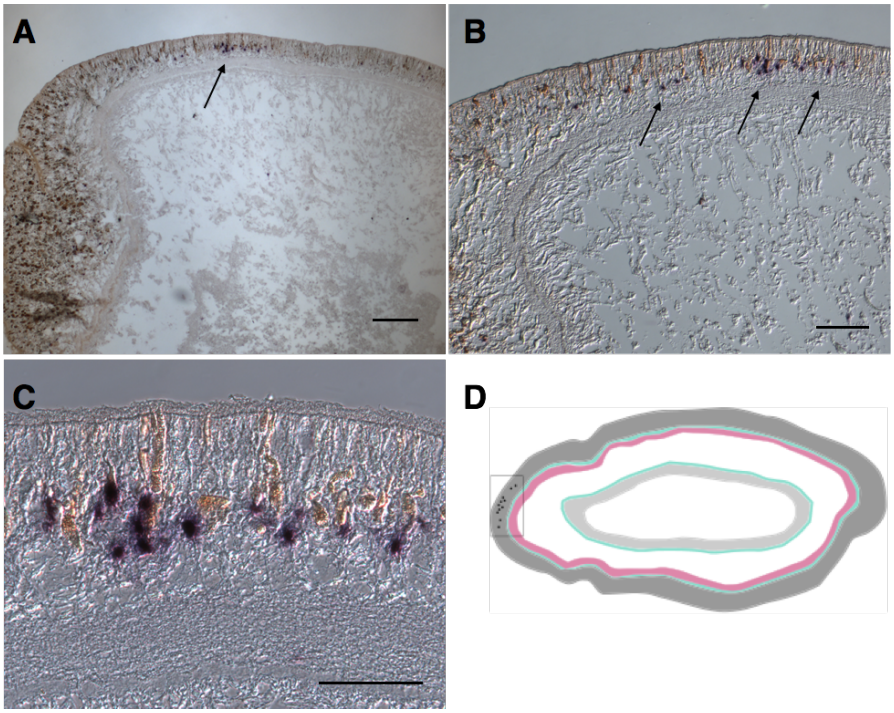


Figure 7.8. *In situ* hybridisation validation of a *Xenoturbella*-specific gene in putative neural cells. Expression in cells found uniquely at the anterior region of the animal. Anterior at the top in both panels. (A) and (B) Overview of expression in cells at the anterior-most region of the animal, location of cells shown by black arrows; (C) Detail of expression in sensory cells embedded in the epidermis; (D) Schematic overview of location (black box) of these cells (represented by black circles) in *Xenoturbella*, anterior to the left of the diagram. Dark grey represents the epidermal layer; two ECM layers shown in green; musculature shown in pink; gastrodermis in light grey. Scale bars: A: 200μm; B: 100μm; C: 50μm.

Overall, *in situ* hybridisation verification provided remarkably strong support for the putative meta-cluster identities based on the single cell libraries.

7.2.7 Meta-cluster specific expression of transcription factors

The identification of putative meta-clusters in *Xenoturbella* can be further investigated by analysing the regulatory mechanisms that underlie common transcriptional states. Whilst comprehensive analysis of differential transcription factor expression across the single cell libraries remains to be carried out, it is clear that some annotated transcription factors are specifically enriched in one or two meta-clusters. For example, *ISL1* – which is known to specify motor neuron identity[193] - is upregulated in the meta-cluster assigned a neural-type identity (Figure 7.9).

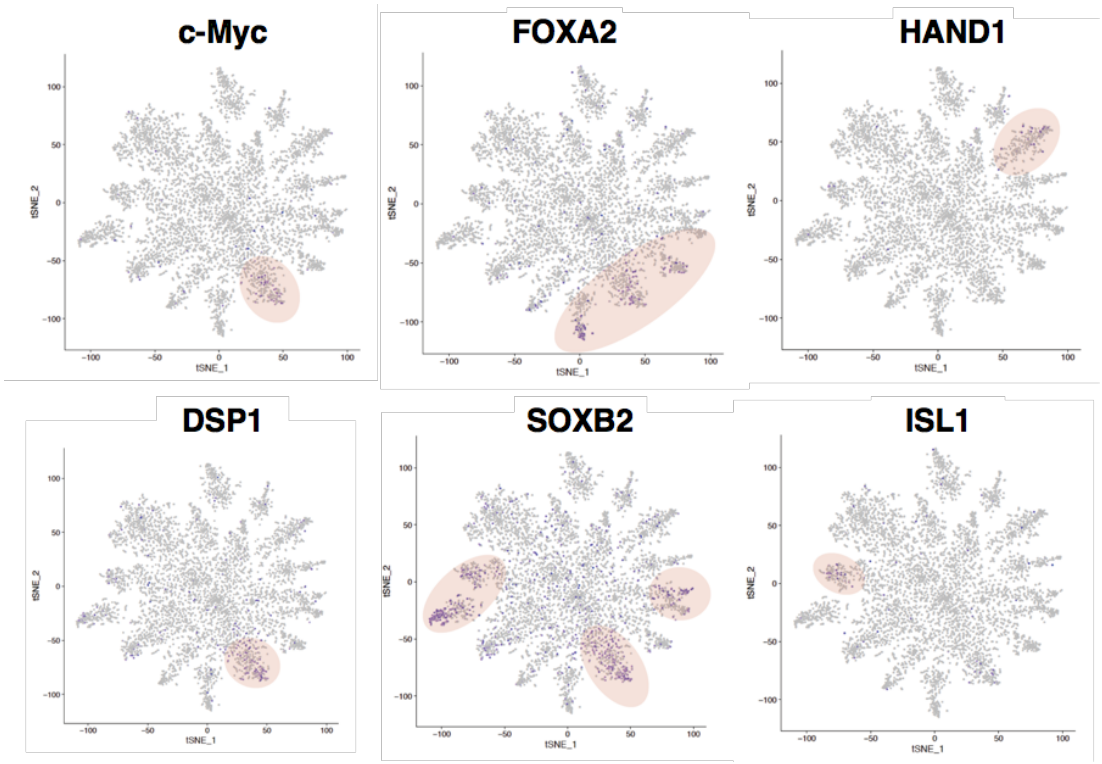


Figure 7.9. Mapping of selected transcription factors onto the Seurat tSNE plot. *c-Myc* (uncharacterised meta-cluster); *FOXA2* (pigmented epithelial); heart and neural crest derivatives-expressed protein 1 (*HAND1*, muscle cells); dorsal switch protein 1 (*DSP1*, uncharacterised meta-cluster); *SOXB2* (sensory epithelial and putative pigmented epithelial); insulin gene enhancer 1 (*ISL1*, neural cells).

7.3 Discussion

7.3.1 Ultrafiltratory-related gene expression

Using the Seurat clustering pipeline, no cell clusters were identified that showed the common upregulation of the five ultrafiltratory genes-of-interest (*XbNeph1, XbNephrin, XbPodocin-like*, *XbCD2AP* and *XbZO-1)*. Across the individual single cell libraries, reads for *XbNeph1* were found in just nine cells, although the number of cells expressing *XbNephrin,* and *XbPodocin-like*, *XbCD2AP* and *XbZO-1* in particular, were more numerous.

Some co-expression of the ultrafiltratory-related genes was found across the single cell libraries. *XbPodocin-like* and *XbCD2AP* are co-expressed in 21 cells; and *XbNephrin* is co-expressed with combinations of *XbPodocin-like*, *XbZO-1* and *XbCD2AP* across the single cell libraries. Although very few cells expressing *XbNeph1* were identified, three of these also express *XbPodocin-like*, and one expresses both *XbPodocin-like* and *XbZO-1.* This is more than would be expected by chance (0.5 cells out of 6008 to co-express *XbNeph1* and *XbPodocin-like* and 0.037 to co-express *XbNeph1*, *XbPodocin-like* and *XbZO*-1). Although cells co-expressing different combinations of these genes were not found in the same cell-type clusters, it is known that *Podocin, ZO-1* and *CD2AP* have roles across multiple tissue types (see section 4.2). Thus, it is likely that cells expressing these genes might be assigned to different meta-clusters based on the up-regulation of other cell-type specific genes.

In situ hybridisation results using probes for *XbPodocin-like*, *XbNeph1* and *XbNephrin* show that the cells of interest in the posterior parenchymal region that appear to express these genes are few in number – although this was variable between sections and individuals (see section 6.3.2). It is therefore likely that the very few cells identified as expressing *XbNeph1* and *XbNephrin* in the single cell libraries could be reflective of the sparse distribution of these cells in the animal.

An alternative explanation for the low number of cells that were identified as expressing *XbNeph1* is the inadequate gene models that are currently in place for *Xenoturbella*. The genome of *Xenoturbella* is not fully annotated, and the 3' preferential sequencing of cells prepared in the MARS-Seq pipeline mean that reads will often be from the 3' UTR region. *XbNeph1* proved particularly problematic to annotate, and so it is possible that more cells do transcribe this gene, but that these have not been identified in the current analysis. As described in Chapter 8, successful mapping to gene models also posed a problem in the *Xenoturbella* Tomoseq protocol, and so further refinement of gene models and 3' extension could help to uncover further reads from these respective libraries.

7.3.2 Meta-cluster identity and diversity

From the Seurat single cell clustering approach, six meta-clusters of cells were identified in the *Xenoturbella bocki* single cell libraries, with identities pertaining to muscle, neuron, gland, epithelium, pigment, and sensory epithelial cell states. The identities of these meta-clusters support what has been proposed by morphological and histological studies of *Xenoturbella*, but the single cell data I present contributes much more than was previously known regarding gene expression and diversity within these cell-types.

7.3.3 Characterisation of clusters using well conserved genes

7.3.3.1 Muscle cells

Of all the meta-clusters, that pertaining to a muscle cell-type identity is particularly well characterised by the presence of genes that are conserved across the Metazoa. Organisation of the musculature in *Xenoturbella bocki* has been confidently described from histological analyses: it possesses a well developed body wall musculature with outer circular muscles; a layer of inner longitudinal muscles lying basally to the circular layer; and radial

muscles that extend from the muscle layer to the gastrodermis (see section 6.1.2, Figure 6.1)[194]. A number of muscle-related genes were found to be enriched in the *Xenoturbella* meta-cluster, including collagen IV, two myosin light chain orthologues, all three troponin orthologues (Troponin C, I, T), Calponin, and a myosin heavy chain orthologue (Figure 7.4). *In situ* hybridisation for two troponin orthologues (Troponin C and Troponin T, for troponin phylogeny see Appendix 8) show strong localised expression in the muscle layer (Figure 7.6). The troponin complex is necessary to mediate contraction in striated muscle in bilaterians; in smooth muscle, contraction is mediated by Calponin. Although no *in situ* hybridisation was carried out with a probe for Calponin, the presence of orthologues for both Calponin and the troponin complex in the muscle meta-cluster indicates that *Xenoturbella* possesses both smooth and striated muscle[195].

7.3.3.2 Neural cells

As found for the muscle meta-cluster, the meta-cluster of neural cells shows upregulated expression of a number of well-conserved neural genes. Synaptotagmin 1-like, Neurotrophin, *ISL1* and a cholinergic receptor, amongst others, are strongly expressed in the putative neural cell cluster (Figures 7.4 and 7.9). These genes are confidently associated with neural cell types in the Bilateria[196-199]. Interestingly, *ISL1* is strongly associated with motor neuron identity and function[193], and its enrichment in a subset of cells within the neural meta-cluster is a possible indicator of the presence of cell types with a motor neuron identity in *Xenoturbella*. Furthermore, a number of genes that are enriched in the neural meta-cluster are also upregulated in the sensory-epithelial meta-cluster (including Elav-like and Synaptotagmin 1-like). This is an interesting finding suggesting a degree of overlap between the nerve plexus and sensory-related cells found in the epidermal layer. Whilst the tSNE plot of Synaptotagmin 1-like expression (Figure 7.4) shows upregulation in the neuronal and sensory epithelial meta-clusters, it is clear that Synaptotagmin 1-like is also enriched in cell clusters across the dataset. Given that Synaptotagmin 1-like is known to function widely in calcium-

mediated vesicle exocytosis, this broad expression pattern is not unexpected[199].

7.3.3.3 Gland cells

Lastly, identity of the meta-clusters representing gland cells and putative pigment cells are also correlated with known cell-type specific genes. Putative gland cells are represented in the tSNE plot as three distinct meta-clusters (Figure 7.3), and mapping gland-specific enriched genes onto the tSNE plot reveals the separation of the gland cell cluster into putative digestive glands (enriched for Chymotrypsin-like and proteases), and another gland cell population characterised by the expression of *Xenoturbella*-specific genes. *In situ* hybridisation using a probe for Chymotrypsin-like revealed specific expression within the gastrodermis, providing support for the identity of digestive gland cells. Subsequent *in situ* hybridisation using a probe for the *Xenoturbella*-specific gene could help identify a novel population of secretory or gland cells with a non-digestive function.

7.3.3.4 Putative pigment cells and phylogeny of Tyrosinase-like genes

The putative pigment cell meta-cluster is characterised by the unique expression of tyrosinase-related orthologues. Tyrosinases are well correlated with pigment cell identity: tyrosinase is necessary for the biosynthesis of melanin, and tyrosinase-like sequences are found throughout the Bilateria and in some diploblasts (cnidarians and sponges) although they are absent in some protostome and deuterostome members (echinoderms, annelids, arthropods)[200]. Tyrosinase sequences have been found to show phylum-specific expansions: phylogenetic analysis of tyrosinases and tyrosinase-like sequences from across the Metazoa has identified four main groups of proteins: cnidarian and protostome tyrosinases; cephalochordate and hemichordate specific tyrosinase-like proteins; chordate 'canonical' tyrosinases; and chordate tyrosinase-related proteins[200]. In assigning identity to the *Xenoturbella* tyrosinase orthologues identified in the pigment meta-

cluster, they were found to group with the cephalochordate and hemichordate specific tyrosinase-like proteins: an interesting finding given the hypotheses regarding the placement of the Xenacoelomorpha as either a basal bilaterian, or as the sister phylum to the Ambulacraria (comprising the hemichordates and echinoderms) (Figure 7.10). *In situ* hybridisation using a probe against one of these Tyrosinase-like sequences confirmed expression in cells found in the epidermis, thought to be pigment cells.

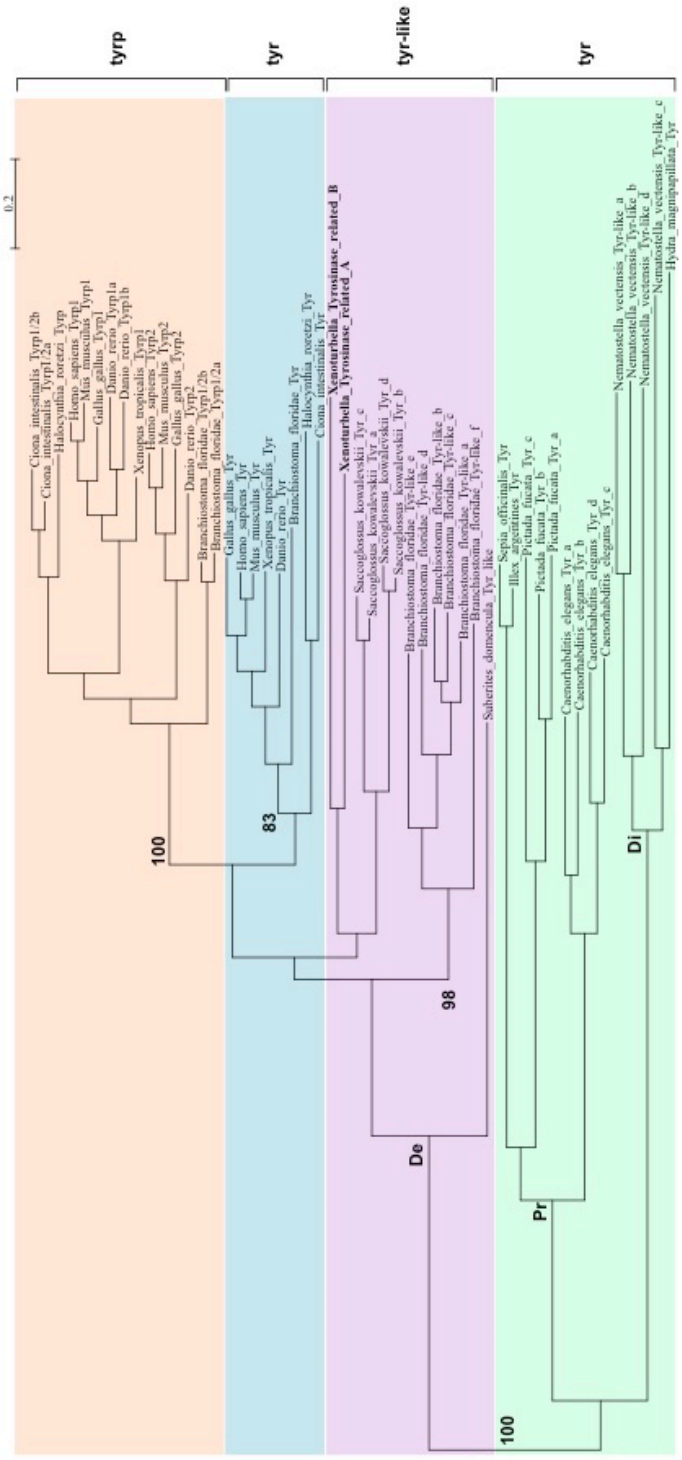


Figure 7.10. Maximum likelihood phylogenetic analysis of tyrosinase-like sequences from across the Metazoa. Cnidarian and protostome tyrosinases (tyr) in green. Cephalochordate and hemichordate specific tyrosinase-like proteins (tyr-like) in purple – including two orthologous sequences identified in the *Xenoturbella bocki* single cell libraries. Chordate canonical tyrosinases in blue; chordate tyrosinase-related proteins (tyrp) in orange. Values at nodes are bootstrap support. Pr = Protostome; Di = Diploblast; Se = Deuterostome. Phylogenetic inference carried out using RAxML. Branch length value shows number of substitutions per site.

7.3.3.5 Sensory epithelial cells and evidence for the presence of primary cilia in Xenoturbella

Interestingly, the meta-cluster identified as sensory-epithelial cells is enriched for the expression of a cilia and flagella- associated gene. Antibody stainings against SrPodocin-like in *Xenoturbella* found strong signal in the ciliated, apical-most portion of the epidermis. In *C. elegans,* the orthologue of Podocin, Mec2, is known to have a mechanosensory role. The identification of cilia-related genes in a putative sensory cell meta-cluster, along with the possible localisation in ciliated cells of a protein with roles in sensory cells in other taxa, could provide a degree of evidence for the presence of both primary and motile cilia in the *Xenoturbella* epidermis. Furthermore, as was previously outlined, some neural genes are also enriched in the sensory epithelial cluster (Synaptotagmin 1-like and *XbElav*), along with a number of *Xenoturbella*-specific genes. This suggests the presence of a *Xenoturbella*-specific group of cells that are involved in neural-like functions but found outside of the main basiepithelial nerve net.

7.3.4 *Xenoturbella*-specific gene expression (*Xenoturbella orphan genes)* in meta-clusters

As well as identifying genes with known functions, the unique transcriptional signatures of each meta-cluster were enriched for *Xenoturbella*-specific genes, for which no similarity could be found with publicly available protein or nucleotide sequences. This was particularly true for the neural cell, sensory epithelial, and putative pigment cell meta-clusters.

The presence of unique, species-specific genes is far from unique to *Xenoturbella.* So-called orphan genes – that is, genes that lack homologues in other lineages, and whose evolutionary origin is frequently poorly understood – have been found to represent up to one-third of the genes in all genomes in all Metazoa[201]. Such genes are hypothesised to arise via a number of different mechanisms, including duplication followed by rapid and

extreme sequence divergence; or de novo emergence, where randomly emerging sequence combinations form functional sites, and co-opt or recruit an enhancer to their regulatory region to become active[201]. In general, genes that are phylogenetically old have been found to have low divergence rates, whereas younger, new genes – such as orphan genes - tend to have faster rates of divergence[202,203]. It is also true that slowly-evolving genes are more highly expressed than younger, fast-evolving genes, and are often implicated in more general functions, or expressed across a broader range of tissues and developmental stages[204].

Most interestingly, *Xenoturbella* orphan genes are amongst those with the most specific meta-cluster enrichment. One of the genes with the most specific neural-cluster expression is a *Xenoturbella*-specific gene (Figure 7.4). *In situ* hybridisation for this uncharacterised gene confirmed a neuronal identity: expression was found in discrete cell bodies above the basiepithelial nerve net at the anterior region of the animal. The expression of a species-specific gene in cells found in the anterior region is an interesting finding, and could suggest a novel sensory function in these cells. Furthermore, *in situ* hybridisation using a probe for *XbElav* also showed enhanced expression in the anterior of the animal (see Figure 6.3). This indicates a degree of concentration at the anterior of the animal of cell types with a neuronal function. As the nervous system of *Xenoturbella* has previously been regarded to lack centralisation or any concentration of nerve cells, this in itself is a valuable finding and would prove interesting to investigate further. In addition, the single cell sequencing libraries indicate much more diversity in neural cell types than was previously thought. The presence of *ISL1* in a subset of neural cells suggests motor neuron specification; and an overlap in expression between neural cells and putative sensory epithelial cells indicates the presence of cells with a sensory-like function outside of the main nerve plexus. *Xenoturbella*-specific genes are particularly enriched in the sensory epithelial meta-cluster, and I hope to localise cells expressing these genes in subsequent *in situ* hybridisation experiments.

Pigment cells in *Xenoturbella* were found to contribute a surprising amount of diversity within the cell-type meta-clusters. Dark pigment spots across the body of *Xenoturbella* are one of their only obvious morphological features, and patterns and colour of pigment spots is variable between individuals (Figure 1.4). The identification of a hemichordate and cephalochordate specific tyrosinase gene in *Xenoturbella* is an interesting finding in itself, but the specific expression of numerous *Xenoturbella* orphan genes in this meta-cluster could also suggest a species-specific function of these pigmented cells.

7.4 General conclusions

Applying whole organism single cell sequencing to *Xenoturbella* is a novel high-throughput approach for uncovering information regarding gene expression and cell-type complexity in this species. Although the single cell data does not help to confidently identify cells that co-express known markers of ultrafiltration, this is not conclusive evidence for their absence. Indeed, the preliminary identification of cells that co-express *XbNephrin* with *XbPodocin-like* and *XbCD2AP* – genes whose protein products are known to interact at the site of ultrafiltration in vertebrates and *D. melanogaster* – is a valuable finding in itself, and lend a degree of support to the *in situ* hybridisation results described in Chapter 6.

The primary transcriptional states identified in this analysis support previous histological and morphological analyses of the animal, but also uncovers a level of interesting diversity within these groupings that was previously unknown. Compared to traditional *in situ* hybridisation approaches, where genes are targeted for investigation without *a priori* knowledge of expression or putative function, single cell sequencing allows for the fast, high-throughput identification of cells with specific molecular markers. Consequently, selection of *in situ* hybridisation probes can be guided to provide a spatial context – and it is evident from the *in situ* hybridisation experiments on *Xenoturbella* that are described in this chapter

that this an efficient and successful approach for the validation of single cell sequencing data.

It is clear that there is still significantly more that can be uncovered from these data. Primarily I aim to re-cluster the meta-clusters outlined in this chapter to further uncover the diversity of gene expression and cell types therein. In particular, the identification of some cell types that are particularly enriched for *Xenoturbella* orphan genes could hint at a novel function, and these would prove interesting for further investigation. In addition, subsequent analyses aim to focus on the diversity of transcription factors in *Xenoturbella*. Although some transcription factors have been identified that show specific meta-cluster enrichment, there is much more that can be done to investigate gene regulation in *Xenoturbella* using the single cell data. In particular, I aim to analyse patterns of transcription factor expression and specificity across the meta-clusters, to investigate the degree of transcriptional hierarchy across different cell types in *Xenoturbella*. Furthermore, some of the meta-clusters in the tSNE plot have not yet been assigned a cell-type or transcriptional state identity. Ribosomal genes are amongst those that are highly expressed in these unassigned groupings, and it is possible that they represent neoblasts or other proliferating cell types, which have been identified and studied in some acoel members[29,177]. Investigating transcription factor specificity and upregulation in cells within these meta-clusters could therefore be informative for understanding proliferative cell types in adult *Xenoturbella* and the genes implemented in differentiation.

8 Tomoseq

8.1 Introduction

8.1.1 Spatially-resolved transcriptomics

As described in Chapter 7, I successfully implemented single cell sequencing in *Xenoturbella* to identify gene expression profiles associated with well-characterised cell types, and to investigate the co-expression of genes that are known to be expressed at the site of ultrafiltration in different bilaterian nephridial systems. However, whole organism single cell sequencing provides no information regarding the spatial distribution of different cell types across the animal. Sequencing individual cells is also not an appropriate technique for identifying genes such as Hox genes that are expressed in multiple cell types but restricted to specific spatial (anteroposterior) domains of expression. Although *in situ* hybridisation and immunohistochemistry can be used to validate cell type results and investigate expression domains – which was used very successfully in *Xenoturbella* - these experiments are limited by the number of genes that can be investigated in parallel. For *Xenoturbella* and other non-model organisms where comparatively little is known about domains of gene expression, this presents a challenge to large-scale gene expression visualisation (*in situ* hybridisation and immunohistochemistry). Even in well-established model organisms, these approaches are not applicable for screening spatial expression of the entire transcriptome.

Combining the manual dissection of tissues or whole organisms with low-input RNA-Seq technologies presents the opportunity for differential transcriptomics with a degree of spatial resolution. This approach was first used in sectioned *D. melanogaster* embryos, but required the use of carrier RNA to overcome problems associated with low starting concentrations of RNA for subsequent cDNA library preparation[205]. The development of RNA amplification protocols optimised for single cell RNA sequencing and/or low-

input starting RNA quantities has negated this dependency on carrier RNA, which can contribute to low numbers of reads-of-interest and background noise. Described by Junker *et al.* (2014)[74], 'Tomoseq' is an RNA tomography approach that combines the manual dissection of tissue (tomography) with RNA-Seq approaches to identify differential gene expression between small tissue sections from across different body axes of an organism.

In the first application of the Tomoseq protocol, zebrafish embryos were cryosectioned into 18μM wide sections across all three body axes (Figure 8.1)[74].

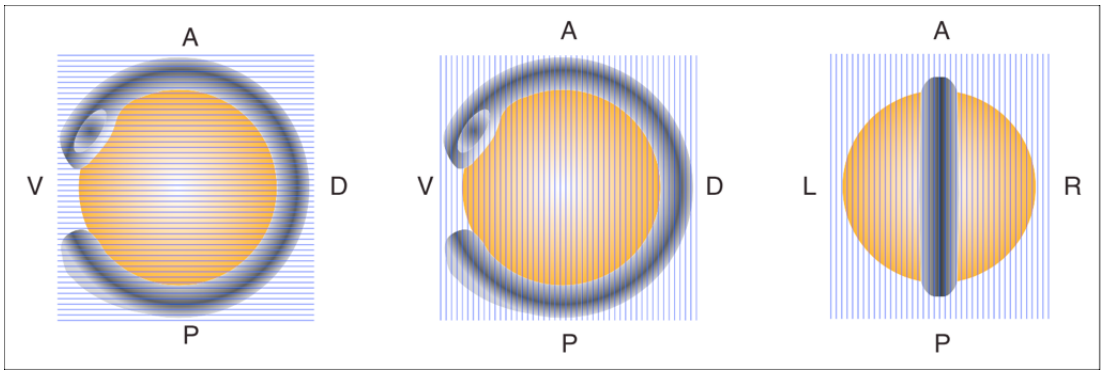


Figure 8.1. Sectioning of zebrafish embryos in the first application of RNA Tomography ('Tomoseq'). Figure taken from Junker *et al.* (2014)[74]. From left to right, panels shows sectioning in an anteroposterior orientation; a dorsoventral orientation; and along the left-right axis of the embryo. A = anterior, P = posterior, D = dorsal, V = ventral, L = left, R = right.

RNA was extracted from the cryosectioned zebrafish embryo tissue and amplified prior to cDNA library preparation. The resulting reads were then mapped to the genome and used to compile a 3D spatial expression pattern of the developing zebrafish on a transcription-wide level. The same approach has also been successfully applied to understand spatial gene expression in regenerating zebrafish hearts[206].

In conjunction with *in situ* hybridisation and immunohistochemistry (Chapter 6) and single cell sequencing approaches (Chapter 7), Tomoseq could be used to investigate variable gene expression along different body axes of *Xenoturbella*. For genes such as Hox genes that are not strongly upregulated in a given cell type, but instead have regionalised domains of expression along a body axis, the single cell clustering approach is not informative for identifying the cells in which these genes are expressed. Furthermore, as outlined, the single cell approach requires *in situ* hybridisation of meta-cluster enriched genes to provide a spatial context for the location of these cells in the animal. Using Tomoseq along different body axes in *Xenoturbella* could therefore offer a complementary approach to the single cell sequencing pipeline, with the ultimate objective of building a 3D map of expression for genes that are upregulated in specific domains of the animal.

8.1.2 Amplifying low starting concentrations of RNA

In order for differential transcriptomics to provide biologically meaningful data, final cDNA libraries must be reflective of the diversity of transcripts represented in the starting pool of RNA. In standard RNA-Seq, there is no pre-amplification of the initial RNA, meaning that preparation of a cDNA library requires a minimal input of at least 1-10ng mRNA[207,208]. This makes standard cDNA library preparation incompatible with sequencing projects such as single cell sequencing and Tomoseq, where such high starting concentrations of RNA cannot be obtained.

For RNA-Seq to be used with very low starting concentrations of input RNA, various approaches for amplifying RNA from tissue sections or single cells have been developed. A number of different amplification-based methodologies have been applied, which use either an exponential PCR-based approach[209,210] or a linear approach[211,212]. For example, in the exponential amplification SmartSeq protocol[209], mRNA is amplified by the ligation of universal primers to either end of the full-length cDNA library,

followed by PCR amplification of all transcripts ('global amplification') using sequences that are complementary to the original universal primers. In linear amplification – for example, that implemented in the CelSeq protocol[211] - a T7 promoter sequence is ligated to the cDNA template, followed by *in vitro* transcription (IVT) by T7 RNA polymerase, which results in >1000-fold amplification of the DNA[211,213].

Although the amplification step is necessary to provide sufficient starting RNA for cDNA libraries, both exponential and linear approaches introduce amplification-dependent noise into the final library when compared to unamplified RNA[214,215]. When comparing transcription levels of genes between RNA samples – as in single cell sequencing and Tomoseq – such amplification biases present a problem for uncovering meaningful differential transcription of genes. The validity of both approaches has been compared in a number of different analyses[214-216], and the merits and drawbacks of exponential vs. linear amplification have been identified in various experiments.

8.1.2.1 Exponential amplification of RNA

PCR-based approaches are recognised to be faster and more cost-effective, providing rapid exponential amplification of RNA[214]. The SmartSeq amplification protocol was originally established using mouse and human cells to improve the yield and length of cDNA libraries generated from individual cells. Although this approach is optimised for a high number of transcripts derived from a small amount of starting RNA, it uses an initial PCR step for exponential amplification of the mRNA, followed by global PCR amplification of all transcripts. Whilst this step helps to improve library yield, it also results in high proportions of primer dimers and spurious PCR products, which contribute to noise in the sequencing data. Although some comparisons indicate that SmartSeq is optimal for amplifying low starting concentrations of RNA[213,216], it has been shown that exponential amplification can be biased towards shorter transcripts with high expression levels[213]. It is

also thought that this approach could systematically reduce gene expression ratios, leading to an overall reduction in the discovery of differentially reduced genes[214].

8.1.2.2 Linear amplification of RNA

Linear amplification and cDNA library preparation – such as CelSeq, and the MARS-Seq pipeline implemented in single cell sequencing (see Chapter 7) avoids the PCR-biases associated with exponential amplification. Instead, the stringent binding of the T7 RNA polymerase to its promoter region means that successful amplification of the RNA pool can be achieved, without spurious PCR products[213,214]. The linear amplification protocol can also incorporate two additional components to act as a control and for later-stage data analysis:

1. The incorporation of a Unique Molecular Identifier ('UMI') into the section-specific barcoding sequences, to eliminate reads in data analysis that might have become duplicated as a protocol artefact (see section 7.2.1).
2. The addition of known quantities of RNA spike-in to each sample prior to amplification, to act as a control and to verify the linear amplification stage.

However, success of the linear amplification IVT step is dependent on the pooling of individually barcoded RNA samples to meet the threshold concentration requirement for T7-mediated linear amplification. There is also the risk that time-dependent RNA degradation during the IVT stage can introduce noise to the final cDNA library[214], and that amplification can be hindered by variable GC content of different sequences[215,216]. Nonetheless, comparisons suggest that linear amplification results in a greater range of transcript lengths, greater estimated mean length, and greater variation of expression levels than exponential amplification[217], and final cDNA libraries

that are more representative of the transcript variability present in the original starting material[215].

8.1.2.3 Implications of amplification biases for the analysis of differential expression

For Tomoseq analyses, changes to the representation of transcripts in the initial starting pool of RNA could mask differential expression of genes with low fold-changes across the body axis. Although for low-input samples, amplifying RNA is a necessary stage in cDNA library preparation, it is not always straightforward to implement the most appropriate strategy to retain meaningful differential expression, especially when the efficacy of amplification may be dependent on characteristics such as nucleotide composition and hairpin structures of the sequences[215].

8.1.3 Tomoseq in *Xenoturbella bocki*

Groups of genes within different classes of taxa can be strongly correlated with discrete domains of expression during development. For example, in the Bilateria, Hox genes and other homeobox-containing genes are well known to be expressed in stereotypical anteroposterior positions relative to one another during development[218]. In chordates, specific domains of expression of several transcription factors are correlated with anteroposterior patterning of the nervous system[183,219]. In the original Tomoseq protocol, differential gene expression was investigated in the developing zebrafish embryo[74]. Whilst it would perhaps be most informative to investigate domains of gene expression in embryonic *Xenoturbella* in order to identify the level of conservation of genes involved in early patterning and development, no embryonic *Xenoturbella* are available for investigation (see 1.4.2). Nonetheless, findings from my single cell data and *in situ* hybridisation results show that some different cell types are associated with specific regions of expression in *Xenoturbella* (see Chapters 6 and 7). Examples of

these include cells with a possible ultrafiltratory function in the posterior parenchymal region; putative neural and sensory cells in the anterior region; and digestive or secretory gland cells in the gastrodermis in the centre of the animal. Investigating how gene expression might differ in RNA extracted from cryosectioned tissue sections across different *Xenoturbella* could therefore provide a degree of spatial context for the cell type meta-clusters identified in the single cell sequencing protocol (Chapter 7).

8.1.4 Objectives of Chapter

I aimed to use the Tomoseq protocol in adult *Xenoturbella* with the objective of investigating the differential expression of genes along the anteroposterior axis. Within this objective, I hoped to investigate differential gene expression from two different perspectives. Firstly, to see if genes with a known bilaterian AP developmental expression pattern were differentially expressed in adult *Xenoturbella*; and secondly, to investigate the expression pattern of ultrafiltratory-related genes and genes identified in the single cell sequencing platform as having a degree of tissue-specific regionalised expression.

Results from the first round of Tomoseq in sectioned *Xenoturbella bocki*, suggested that biases associated with PCR-based amplification could be a source of experimental error, perhaps masking any differential expression of genes along the anteroposterior body axis. Consequently, the second objective of this chapter was to establish the CelSeq2[212] RNA linear amplification and cDNA library preparation protocol as part of the Tomoseq assay, in order to investigate biologically meaningful differential expression across the *Xenoturbella* AP axis. The linear amplification of RNA extracted from small tissue sections is a novel technique to establish in *Xenoturbella*.

8.2 Results

8.2.1 Initial Tomoseq results using non-linear amplification of RNA

An initial Tomoseq assay was carried out by sectioning an adult *Xenoturbella* into 96 pieces along the anteroposterior axis. After my extensive refinements of the RNA extraction protocol for low quantities of starting tissue (see section 2.11.2), RNA from these sections was sent for amplification and library prep using the SmartSeq2 exponential RNA amplification protocol, and sequencing.

A total of ~90,000 contigs from HiSeq sequencing were assembled using the CLC assembly cell. Of these contigs, approximately one-third (~28,000) mapped to bacterial sequences when blasted against the UniProt database. The remaining ~60,000 contigs were quantified using kallisto and differential expression across the 96 sections analysed using sleuth (for detail see section 2.11.5.1). Heat maps across the 96 sections were generated for Hox genes (*Hox1*, *Hox4*, *Hox7* and *Hox11*) and anteroposterior (AP) neural patterning genes[183] for which orthologous sequences could be identified in the *Xenoturbella* transcriptome (*Elav*, *Musahashi*, *Sox1/2/3, Six3*, *Bf1*, *Dlx*, *nk2.1*, *Pax6*, *Tll*, *barH*, *Otx*, *Lim1/5*, *Gbx*, *Emx*, *Dbx*, *Vax*, *Rx*, *Irx* and *Engrailed*) (Figure 8.2). It is clear from these heat maps that no differential expression across the AP axis is seen in the RNA-Seq data for these genes: expression is absent for many AP neural patterning genes across many sections, apart from relatively high expression seen ubiquitously for *Musahashi*, *Sox1/2/3* and *Six3*. Similarly, Hox gene orthologues have no detected expression across most sections, with peaks of expression for some sequences across the sections not corresponding to an anteroposterior pattern.

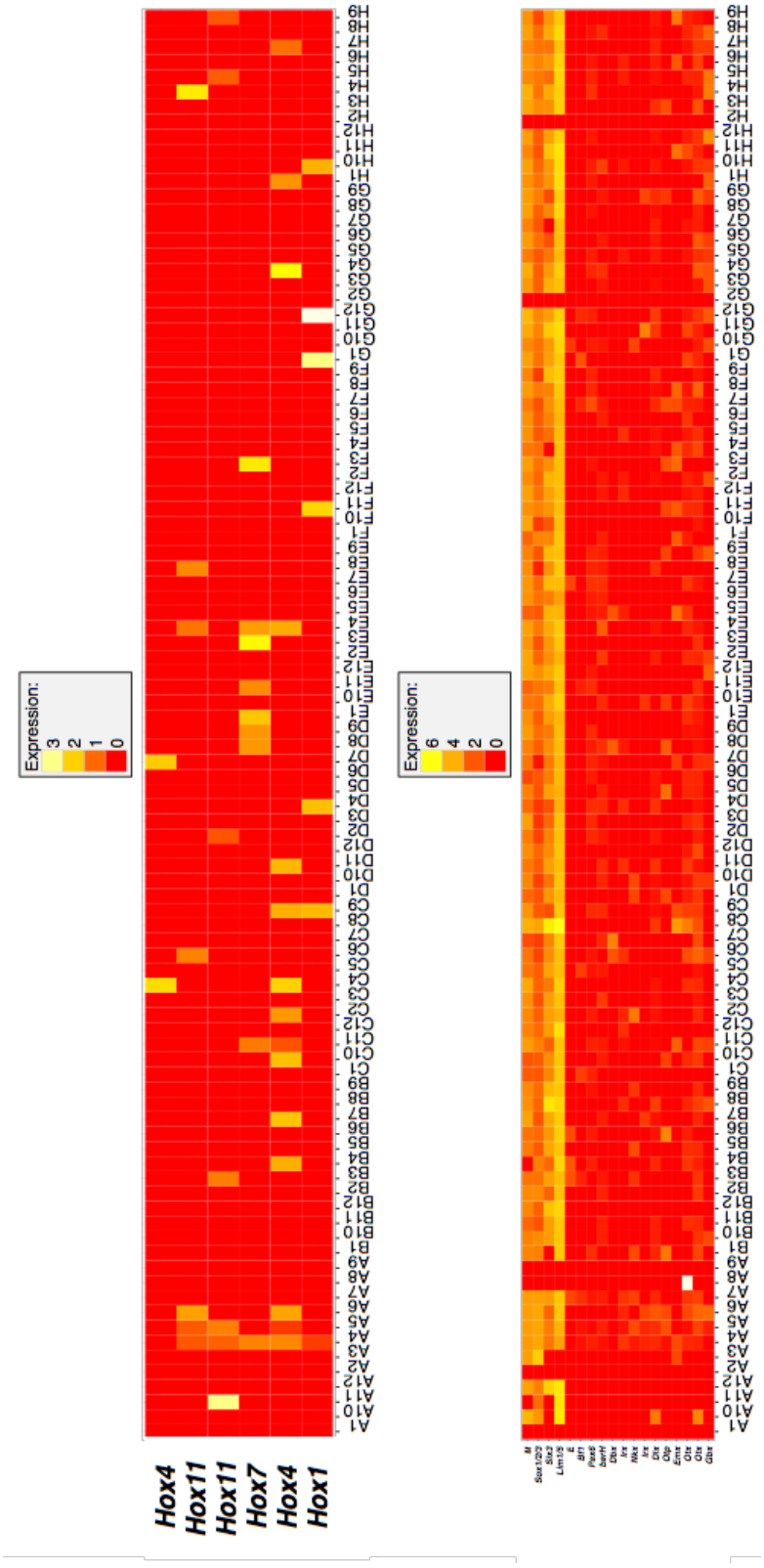


Figure 8.2. Heat maps showing relative levels of expression for selected genes in *Xenoturbella* from RNA libraries prepared using the SmartSeq2 protocol. Heat map from anterior-most (left) to posterior-most (right) sections (A1 to H12). (A) Hox genes (B) Chordate AP patterning genes[183]

A principal component analysis (PCA) also indicated that there was no structure in differential gene expression across the AP axis in libraries prepared using the SmartSeq2 protocol (Figure 8.3).

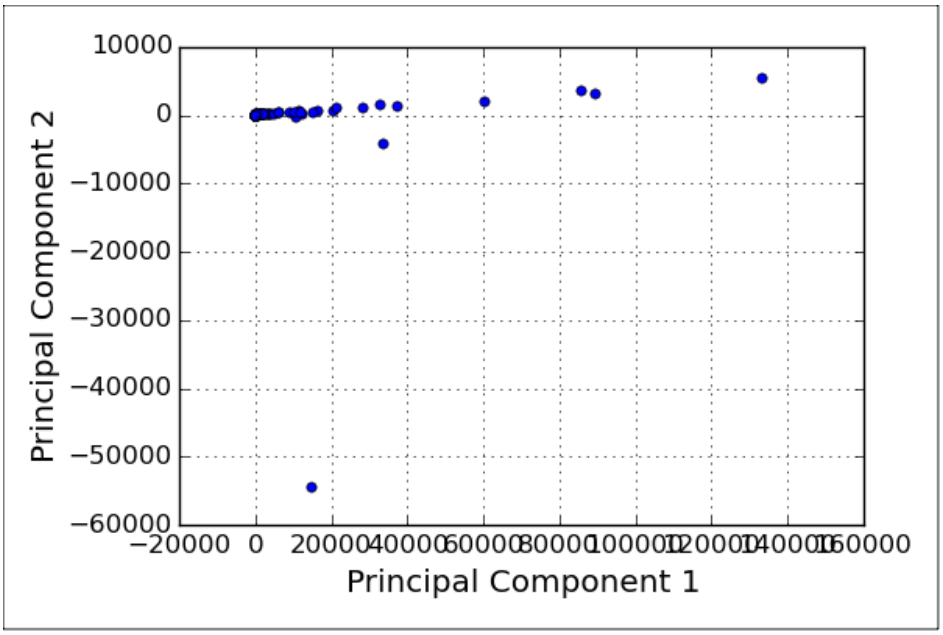


Figure 8.3. PCA of transcriptomic data across 96 anteroposterior sections, from RNA libraries prepared using the SmartSeq2 protocol. Most sections seem to cluster at the same point, indicating no variance in transcriptomic signature across the different sections. PCA carried out by Philipp Schiffer (Telford lab).

8.2.2 CelSeq2 linear amplification in *Xenoturbella bocki*

8.2.2.1 Read numbers and mappings across the libraries

Given the lack of any differential expression from the initial Tomoseq assay, a second animal was cryosectioned along the anteroposterior axis into 90 ~60μM sections. RNA was extracted from each section using the same RNA extraction protocol as was used for the first round of Tomoseq, described in detail in section 2.11.2. To see if exponential amplification of RNA was masking any differential expression of genes across the AP axis, cDNA libraries for this second Tomoseq assay were prepared using the CelSeq2 linear amplification protocol.

As part of the CelSeq2 protocol, individual RNA samples were labelled with one of 20 differently barcoded primers, which included a unique molecular identifier sequence (UMI), to allow for identification of any PCR duplicate sequences. A defined amount of ERCC synthetic spike-in was also added to each sample to verify linear amplification of the RNA pool. cDNA was prepared as five final libraries, with each library differentially labelled with a different Illumina RPIX sequences (Figure 8.4). All five libraries were pooled at equimolar concentration for HiSeq sequencing. For a detailed protocol, see section 2.11.3.

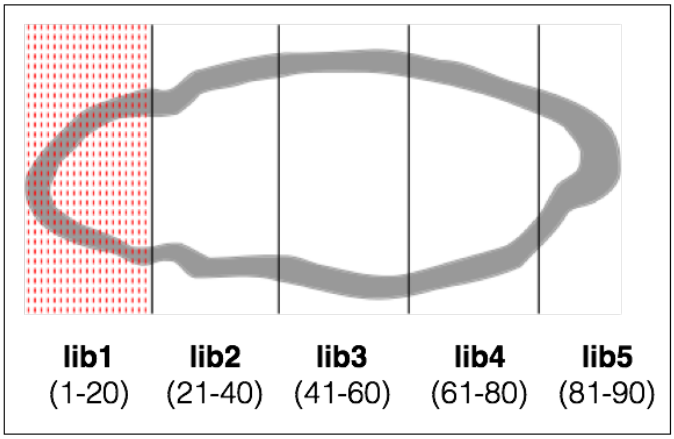


Figure 8.4. Schematic representation of labelling and pooling of sections across the AP axis of *Xenoturbella*. Anterior to the left of the diagram. Sections representing libraries 1-5 labelled as lib1-5, demarcated by solid black lines. Red dashed lines in library 1 represent the 20 sections (10 in library 5) comprising each library.

RNA-Seq (Illumina Hi-Seq) of the final pooled libraries prepared using CelSeq2 yielded a total of ~180 million reads (100 bp paired-end) from the cryosectioned *Xenoturbella*: a 1500x increase compared to the number of reads that were obtained in the first Tomoseq assay (Figure 8.5A). To test for linear amplification of the ERCC spike-ins, all reads per library were mapped against the set of spike-ins using RSEM. TPM values from RSEM were plotted against the starting concentration using sleuth in R. Spike in amplification showed a linear response across all sections, with a Spearman's correlation (rho) of 0.87 ± 0.04 (Figure 8.5B).

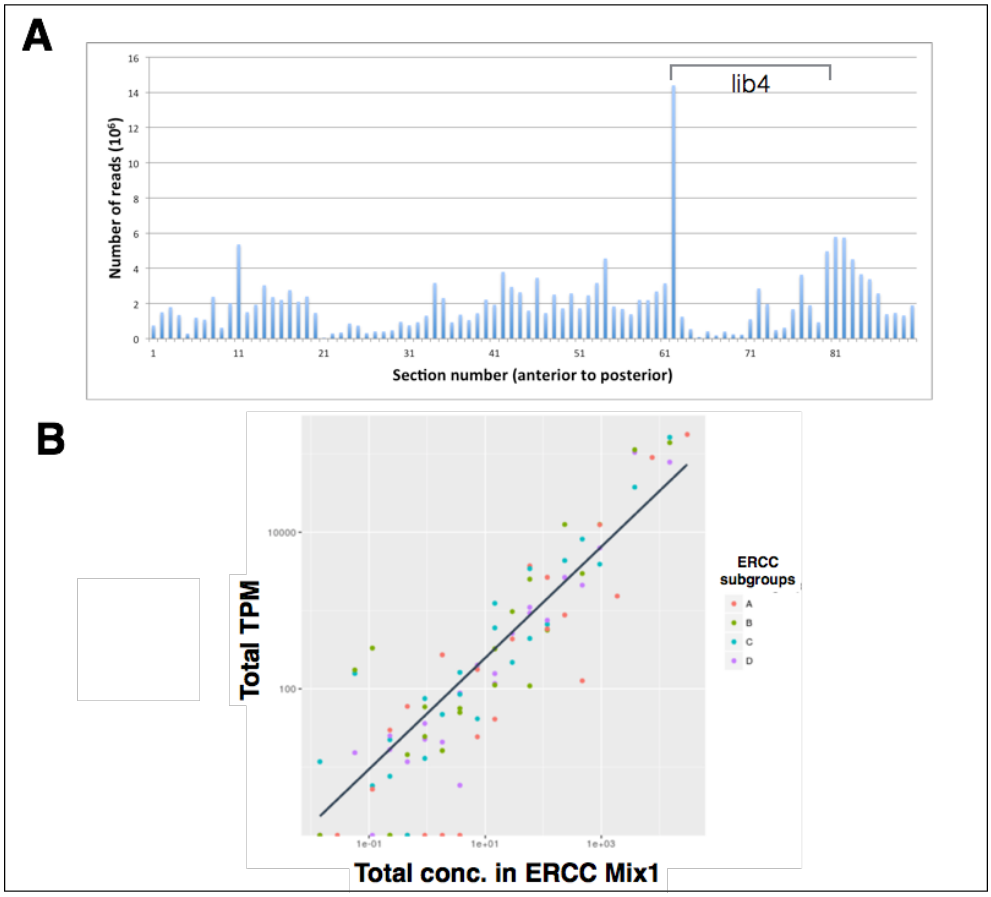


Figure 8.5. Success of the CelSeq2 protocol in *Xenoturbella*. (A) Total number of reads (10^6) per section. Problematic sections from library 4 indicated by grey bracket (lib4). (B) CelSeq2 protocol on RNA extracted from *Xenoturbella* shows a linear response across all sections (shown here in library 1). The plot shows normalised expression of the four ERCC spike-in groups, with an idealised linear amplification shown by the black line. Analysis carried out in conjunction with Philipp Schiffer (Telford lab).

In library 4, which included all sections between 61 and 80, an error in amplification or library preparation resulted in far fewer reads than were found in the other sections (Figure 8.5A). An anomalously high number of reads was found in the second tissue section included in this library (cryosection 62), comprising over 14,000,000 reads compared to a mean of nearly 2,000,000 across the rest of the sections. Followed by this peak, sections 63 to 70 have far fewer reads than the rest of the data set, with a mean of just 421,000 reads per section. Although the number of reads from sections 71-80 are higher, and have a mean of ~2,000,000 in line with the rest of the sections, the percentage of reads that map to the *Xenoturbella*

genome from this half of the library are very low (8-30%), indicating a protocol error.

Of the total reads from across all sections, approximately 70-90% mapped to the *Xenoturbella* genome. Between 7-24% of the reads per section mapped to the ERCC spike-in sequences. Of the remaining reads, a maximum of 25% mapped to bacterial sequences – with a mean across all sections of 15.04% - composed predominantly of Proteobacteria sequences, and a small percentage of Chlamydia sequences, which is a known *Xenoturbella* symbiont[220].

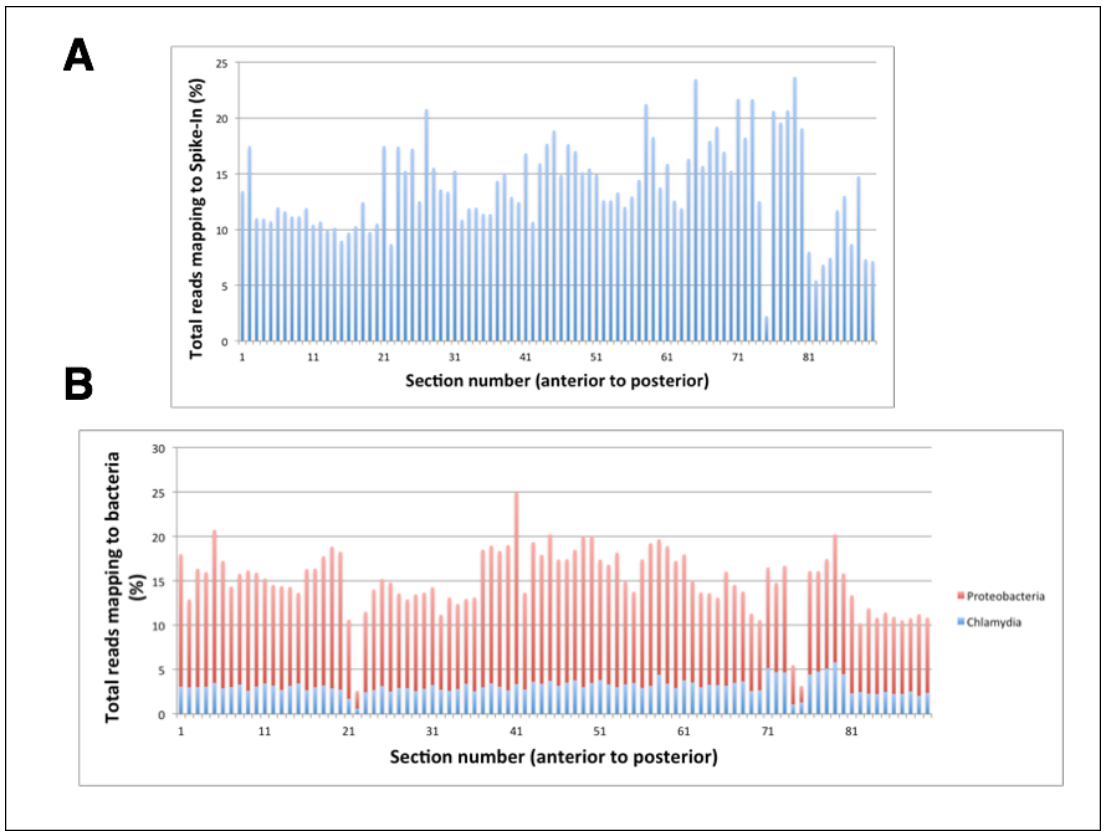


Figure 8.6. Spike in and bacterial mapping from reads using CelSeq2 in *Xenoturbella*. (A) % of reads per section mapping to ERCC spike in; (B) total % of reads per section mapping to bacteria (Proteobacteria in red; Chlamydia in blue). Analysis carried out in conjunction with Philipp Schiffer (Telford lab).

Total reads mapping to the *Xenoturbella* genome were deduplicated based on UMI sequence using umi_tools (https://github.com/CGATOxford/UMI-tools) (see section 8.1.2). Total UMI count across all 90 sections was ~18,000,000 with an average UMI count of ~203,000 UMIs per section. It is evident from UMI distribution across the data set that library 4 (comprising sections 61-80) has a much lower UMI count, as would be expected given the smaller number of reads contributed by this library (Figure 8.7). When the anomalously high UMI count contributed by section 62 is excluded from analysis, mean average UMI count in library 4 is ~62,500 per section.

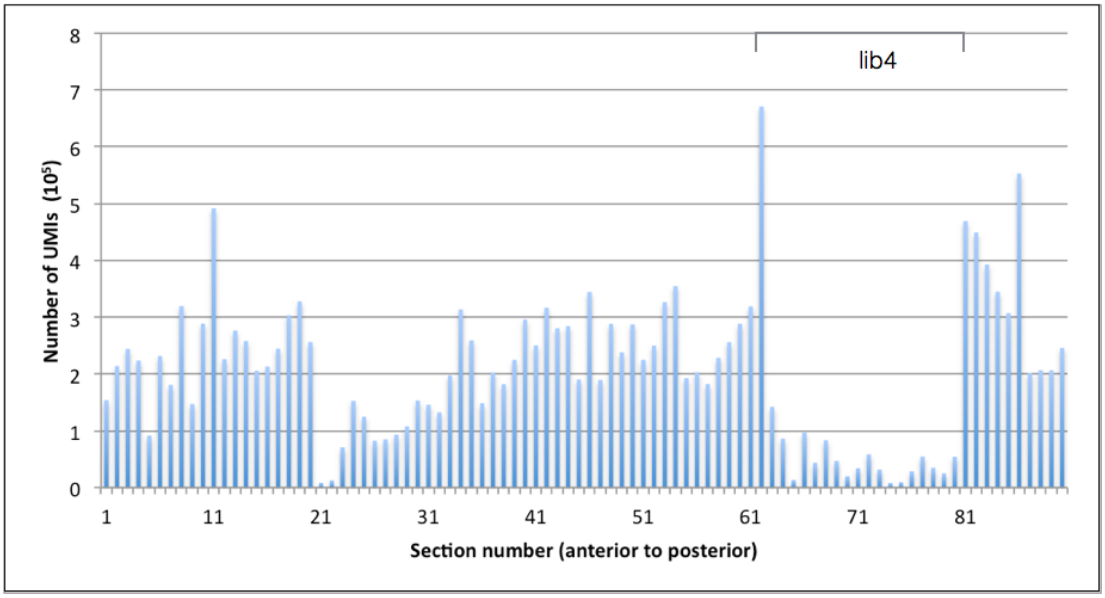


Figure 8.7. Total number of UMIs (10^5) per tissue section. Problematic sections from library 4 indicated by grey bracket (lib4). Analysis carried out in conjunction with Philipp Schiffer (Telford lab).

8.2.2.2 Clustering the data set

To see if RNA-Seq data from *Xenoturbella* sections generated as a result of linear amplification in the CelSeq2 protocol showed any overall AP trends, PCA was carried out. RNA-Seq data from each of the 90 sections were grouped as 9x 10 sections to represent three anterior, three mid-section, and three posterior regions across the AP axis of the animal (Figure 8.8). The second posterior section (post2), represented by sections 71-80, falls entirely within the problematic library 4. It is clear from the PCA plot that this half of library 4 constitutes the second principal component of the dataset. A lot of the variance seen in the data can therefore be accounted for by differences in the transcriptomic signature of this library in comparison to the rest of the data set. This is a likely indicator of a contamination or protocol error. The first posterior section (post1), representing sections 61-70, and also from library 4, has a much broader distribution than the other groups of sections.

Nonetheless, unlike the PCA plot representing the SmartSeq2-amplified RNA, it appears that there is clear structure in the data, with anterior and posterior sections distributed differently in the PCA compared to those originating form the middle of *Xenoturbella* (Figure 8.8). We can see patterns in the overlapping clusters of sections: the three mid-section clusters overlap very closely; the first two anterior regions overlap with one another and with the third posterior region, and the third anterior region overlaps with the first posterior region. Furthermore, these similarities in transcriptomic signature can be interpreted in line with what we know of *Xenoturbella* morphology and the cell types identified in the single cell sequencing analysis (see 7.2.4). In particular, the close overlap of all sections defined as the mid-section could reflect the common upregulation of genes associated with gut and the proportional reduction of other genes in this region.

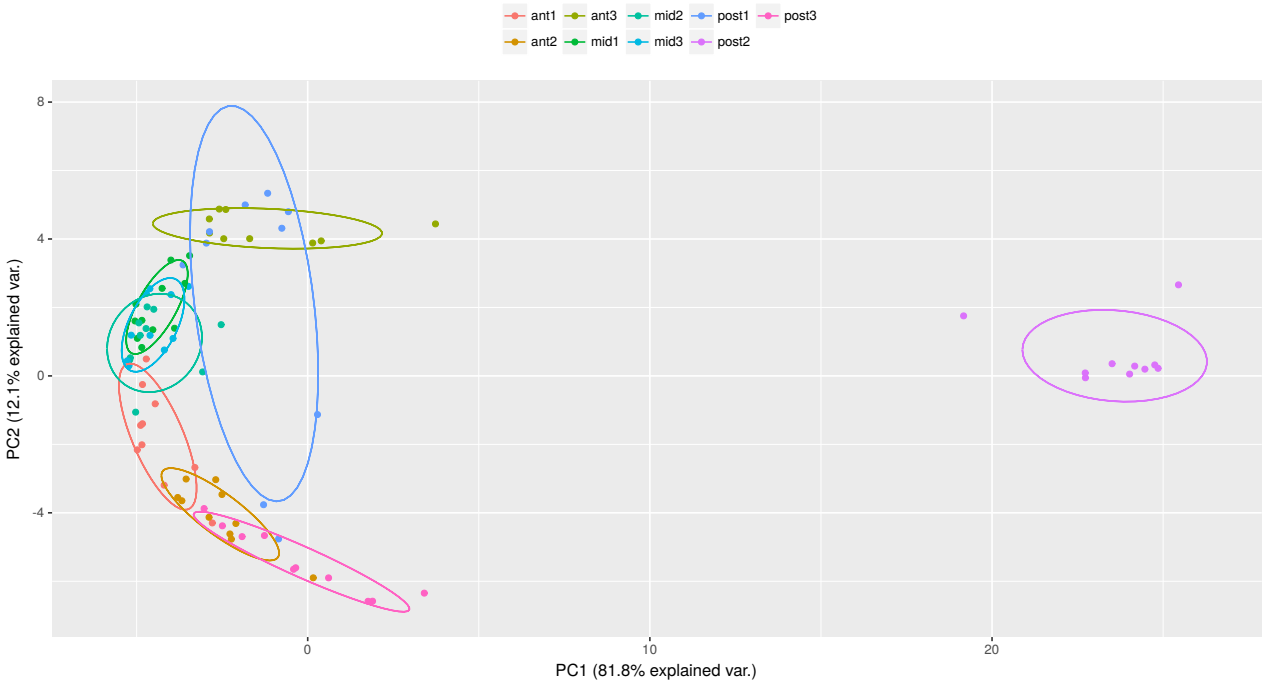


Figure 8.8. PCA of Tomoseq sequencing data from *Xenoturbella* RNA amplified using the CelSeq2 protocol. Sections along the anteroposterior axis grouped into blocks of 10: 3x anterior (ant1, ant2, ant3); 3x mid-section (mid1, mid2, mid3); and 3x posterior (post1, post2, post3). PCA analysis carried out by Philipp Schiffer (Telford lab).

8.2.2.3 Distribution of common bilaterian AP developmental genes across Xenoturbella

As investigated in the first Tomoseq assay using SmartSeq2 PCR based amplification, heat maps across the 90 sections were generated for Hox gene orthologous sequences and the same chordate AP patterning genes[183] (Figure 8.9). Much like the results seen in the first Tomoseq data, these heat maps show patchy expression for all genes investigated, which is not correlated with differential expression along the AP axis. Many of the genes also have zero coverage across all sections, which is indicative of 'missing' hits against current gene models.

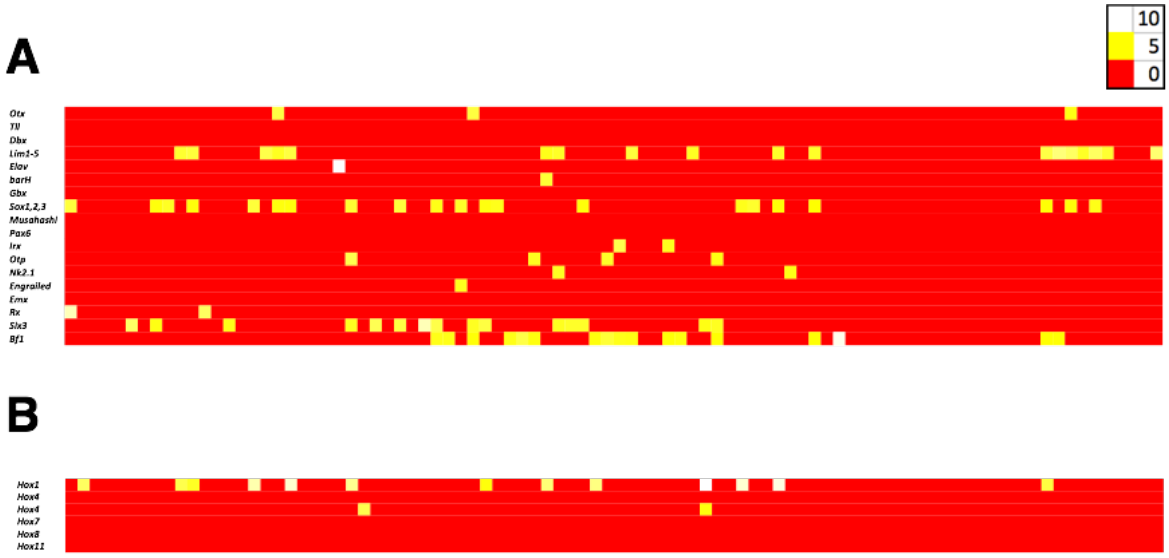


Figure 8.9. Heat maps showing relative levels of expression for selected genes in *Xenoturbella bocki*. Heat maps from anterior-most (left) to posterior-most (right) sections, generated from RNA linearly amplified using CelSeq2. Expression levels are FPKM. (A) Chordate developmental AP neural patterning genes[183] (B) Hox genes.

A lack of meaningful signal in adult *Xenoturbella* from genes with a stereotypical AP expression in developing bilaterians does not mean that the data lack any AP variance. As evidenced by the PCA, the transcriptional signatures across different domains of the animal do appear to show region-specific identity. An absence of differential AP expression for these genes could be attributed to the use of an adult animal in Tomoseq, as opposed to an embryonic animal. Consequently, genes that were identified from single cell sequencing analysis to have a specific cell-type expression in the adult animal were investigated for AP differential expression in the CelSeq2 data.

8.2.2.4 Cell-type specific expression across the AP axis

Of the six meta-clusters of putative cell types identified from *Xenoturbella* whole organism single cell sequencing (see section 7.2.4), three were targeted for AP expression analysis: neural, gland and muscle. In the single cell data, cells in the neural and gland cell meta-clusters were found to have transcriptional signatures characterised by the up-regulation of a number of annotated and *Xenoturbella*-specific genes. The distribution of cells that are up-regulated for these genes was verified by *in situ* hybridisation, and showed the patchy distribution of neural and gland cell types, located in greater density in the anterior and gastrodermal regions of the animal, respectively (see section 7.2.6). Muscle cells are upregulated for many known muscle genes, including troponin and myosin members. Distribution of muscle cells is more ubiquitous across the animal, with muscle cells found in a defined layer underlying the epidermis across all body axes (see section 7.2.6). Genes associated with these neural, gland and muscle cells were therefore targeted for investigation in the Tomoseq CelSeq2 data, with their known regional (neural and gland cells) or ubiquitous (muscle) expression across the animal.

It is clear from the heat maps generated (Figure 8.10) that the sections covered by library 4 (61 to 80) cannot be used to infer patterns of gene expression in this region: FPKM expression for all genes analysed in

these heat maps (with the exception of one unannotated gene in the muscle gene heat map) is 0.

Heat maps generated for neural and gland cell related gene expression over the 90 sections show a sporadic distribution across the AP axis. Whilst single cell data, confirmed by *in situ* hybridisation, show that genes expressed in the neural cell cluster localise to cells with enhanced anterior distribution, the neural heat map does not show this (Figure 8.10A). An uncharacterised and an unannotated gene (see section 7.2.4) show broad expression across the AP heat map, as does a neuron-specific kinase activator called *CDK5R1*. However, these genes also show enhanced expression across different cell clusters in the single cell library data, and so broader expression across the AP axis would perhaps be expected. For other neural-related genes, expression is patchy but not strongly correlated with upregulation in the anterior region, as might have been anticipated. No gland-cell related genes have a broad AP distribution: instead, they are found upregulated in individual, apparently randomly distribute cryosections, with the exception of a Chymotrypsin-like sequence, which is found upregulated in four clusters of successive sections (Figure 8.10B). Whilst these domains of expression do not reflect the findings of the single cell library data or *in situ* hybridisation, it is possible that 'missing' hits for transcripts from the CelSeq2 data are likely hindering interpretation of gene distribution: for example, Synaptotagmin 1-like, and various proteases, which were confidently identified in the single cell data, have FPKM values of 0 in the respective neural and gland cell heat maps.

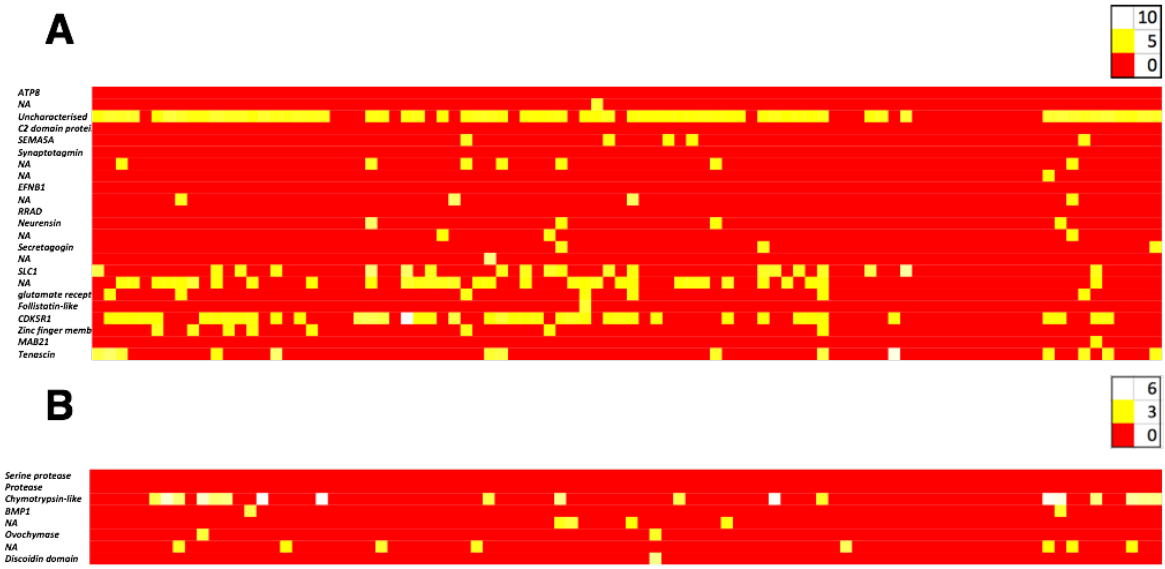


Figure 8.10. Heat maps showing relative levels of expression for selected neural and gland-related genes across the AP axis of *Xenoturbella.* Heat map from anterior-most (left) to posterior-most (right) section, from RNA linearly amplified using CelSeq2. Expression levels are FPKM. (A) Neural-cell related genes; (B) Gland-cell related genes.

Similarly, in the muscle gene heat map (Figure 8.11), expression for all three troponin genes is 0, despite *in situ* hybridisation confirming the expression of both Troponin T and Troponin C very specifically in the *Xenoturbella* muscle cells (see Figure 7.6). It is possible that the preferential 3' sequencing of the RNA libraries amplified using CelSeq2 means that not all transcripts have been assigned to the correct gene model, and have thus not been included in analysis. Nonetheless, for other muscle-related genes, expression is broader than was found for the neural and gland-cell genes (Figure 8.11). This is particularly true for both myosin light and heavy chain. Given the presence of the muscle layer across the AP body plan of *Xenoturbella*, this is an expression pattern that might be expected.

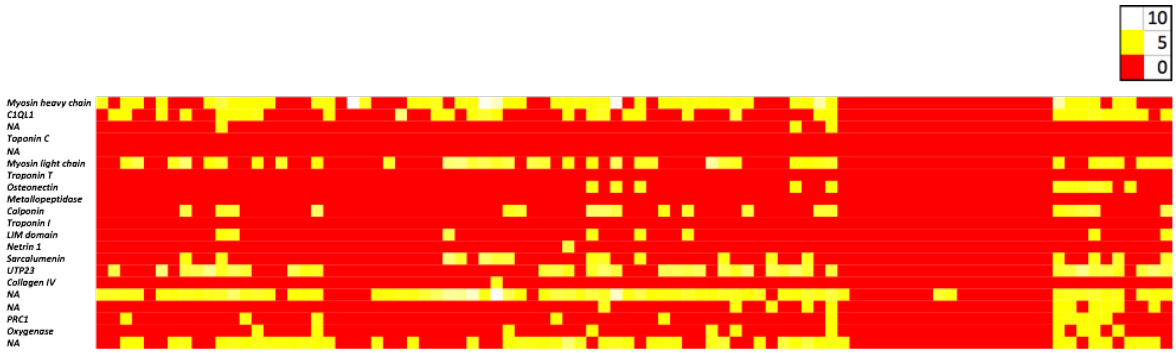


Figure 8.11. Heat map showing relative levels of expression for muscle-related genes across the AP axis of *Xenoturbella.* Heat map from anterior-most (left) to posterior-most (right) section, from RNA linearly amplified using CelSeq2. Expression levels are FPKM.

8.2.2.5 Expression of ultrafiltratory-related genes

Having identified cells using *in situ* hybridisation in the posterior region of the animal that appear to express three genes related to ultrafiltration (*XbNeph1*, *XbNephrin* and *XbPodocin-like*) (see Chapter 6), I also wanted to investigate the expression of these genes in the CelSeq2 data. Two other genes are also known to function at the site of ultrafiltration in vertebrates and *D. melanogaster* (CD2AP and ZO-1), and so sequencing data from *Xenoturbella* orthologues of these genes were also included in analysis (see sections 4.1.3 and 4.2).

It is possible that the 3' preferential sequencing of libraries prepared using the CelSeq2 protocol has resulted in unassigned transcripts for *XbNeph1* and *XbNephrin* in the Tomoseq data (Figure 8.12). RNA probes for both genes confirmed their expression in *in situ* hybridisation, and single cell libraries also identified the expression of these genes (albeit in relatively few cells for *XbNeph1*). Consequently, the 0 coverage found for *XbNeph1* across all sections, and the very sporadic expression of *XbNephrin,* is likely an artefact.

However, the single cell sequencing libraries also show that cells expressing *XbPodocin-like*, *XbCD2AP* and *Xb-ZO-1* are more numerous than those that were found to express *XbNeph1* and *XbNephrin*. The heat map for these genes appears to correspond with this finding: *XbPodocin-like* and *XbCD2AP* have upregulated expression in groups of sections across the AP axis, and *XbZO-1* is also found in patchy individual sections across the animal (Figure 8.12). Furthermore, some sections show the common upregulation of more than one of these ultrafiltratory genes: a block of six successive sections at the posterior of the animal has enhanced expression for both *XbCD2AP* and *XbPodocin-like*, which are known to interact at the site of ultrafiltration in both the vertebrate podocyte and insect nephrocyte (see section 4.1.3).

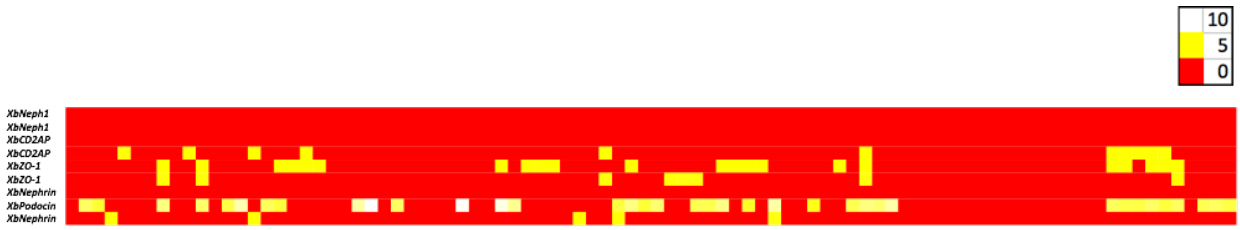


Figure 8.12. Heat map showing relative levels of expression for ultrafiltratory-related genes across the AP axis of *Xenoturbella*. Heat map from anterior-most (left) to posterior-most (right) section, from RNA linearly amplified using CelSeq2. Expression levels are FPKM.

8.3 Discussion

8.3.1 Refining RNA extraction from small tissue samples of *Xenoturbella*

Extracting RNA from cryosectioned *Xenoturbella* tissue samples required a great deal of refinement prior to subsequent amplification and library prep. Various parameters that I modified including the snap-freezing of tissue sections immediately after cryosectioning; placing tissue samples directly into frozen or room temperature Trizol; ribolysing or vortexing the tissue; and including RNase inhibitor in the final eluted RNA. Ribolysing the tissue in particular was found to significantly hinder RNA extraction, likely owing to shearing of the RNA. Of all parameter combinations, optimal RNA extraction was achieved from samples placed into room temperature Trizol, vortexed briefly, and stored at -80°C. RNase Inhibitor was included in the final RNA elution (for full protocol see section 2.11.2). Using this protocol, concentrations of between ~2ng/µl and ~70ng/µl RNA (variable depending on the cross section of the region of the animal that each tissue section corresponded to) were successfully extracted from the cryosectioned tissue, all of which resulted in two visible bands of rRNA on ethidium-bromide stained gels, as would be expected for undegraded RNA samples. This indicated that my final RNA extraction protocol for *Xenoturbella* was reliable and applicable to small amounts of starting tissue.

8.3.2 Non-linear RNA amplification: implications for spatial transcriptomics

RNA extracted from the first round of Tomoseq was amplified externally using the SmartSeq2 protocol. A PCA showed that there was no biologically meaningful structure present in the data generated from across the anteroposterior axis. However, *in situ* hybridisation experiments on tissue sections of the adult animal confirm the spatially constrained expression of certain genes (see, for example Figure 7.7 and 7.8) in cell types with a

specific function. With this in mind, it was likely that the lack of any identifiable change in the expression of genes along the AP axis could be attributable to experimental error or artefacts, rather than being a biologically true result.

Three possible sources of experimental artefact were identified in the initial *Xenoturbella* Tomoseq protocol:

1. Transcript bias in the RNA extraction protocol, masking any biologically real differential expression.
2. Tissue residue remaining on the blade of the cryostat and causing contamination between successive sections.
3. Differential transcript bias in the RNA amplification protocol, causing enhanced amplification of some transcripts over others to mask any true differential expression.

Having refined a reliable RNA extraction protocol from small starting quantities of tissue in *Xenoturbella*, changing the approach for RNA extraction was not a favoured strategy. Furthermore, very little literature is available regarding bias in RNA extraction, perhaps indicating that this potential source of error was not as relevant. Preliminary experiments where the blade was cleaned or moved between sections changed the angle of the blade, causing an inconsistent cutting plane and different widths of tissues in successive sections. Whilst it is possible that there would have been some cross-contamination of tissue between sections, this is unlikely to have been sufficient to mask differential gene expression across the whole animal. Maintaining a constant cutting angle to ensure uniform tissue width across the sections was therefore prioritised over adjusting or cleaning the blade to prevent tissue crossover.

Of all sources of experimental artefact, that of amplification bias introduced through non-linear PCR RNA amplification was targeted for protocol refinement. Although the SmartSeq2 protocol is optimised for a high number of transcripts derived from a small amount of starting RNA, it uses a

PCR step for exponential amplification of the mRNA, followed by global PCR amplification of all transcripts[90]. Whilst this step helps to improve library yield, it can also result in high proportions of primer dimers and introduce a number of PCR biases. In particular, SmartSeq PCR biases have been reported to over-amplify transcripts with high expression levels and short lengths[213]. It is possible that such biases would mask differential expression of genes with low fold changes across the *Xenoturbella* anteroposterior axis, contributing to the lack of any variance observed in the initial Tomoseq trial.

To properly establish Tomoseq in *Xenoturbella* and investigate biologically meaningful differential gene expression across the anteroposterior axis, I chose to implement the CelSeq2 linear amplification protocol in *Xenoturbella*[212].

8.3.3 Establishing linear amplification and RNA library preparation

Originally designed for amplification of RNA from for single cells, I aimed to optimise the CelSeq2 protocol for clean, low-concentration *Xenoturbella* RNA samples.

Based on the most successful combinations of parameters optimised during the 20-section protocol refinement (see section 2.11), I used the CelSeq2 protocol on RNA extracted from 90 sections of *Xenoturbella* (see 2.11.3). Using the ERCC Spike-In as a control shows that linear amplification of RNA had been successfully implemented in *Xenoturbella*: a novel approach that had not been previously established in this species.

8.3.4 Tomoseq with linear amplification (CelSeq2)

8.3.4.1 Data quality

RNA-Seq data from libraries 1, 2, 3 and 5 prepared in the CelSeq2 pipeline show a high proportion of reads mapping to the genome, at between 70% and 90% of the overall reads obtained. It is clear that a protocol error occurred during the amplification and preparation of library 4. A possible explanation for this is primer contamination between the primers used in section 62 and the following 8 primers. Consequently, reads from sections 63 to 70 may have been assigned to the barcode found in the primer used in section 62, causing the abnormally high number of reads in section 62, and the depletion of reads in the rest of the first half of this library. As >70% of the reads in section 62 still map to the *Xenoturbella* genome, this appears to be a likely explanation for the erroneous coverage found in the first half of library 4. Nonetheless, the percentage of reads mapping to the genome across the second half of library 4 is between 8% and 30%, which could instead suggest a contamination error for these sections.

In the initial Tomoseq assay, over one third of the total assembled contigs mapped to bacterial sequences. In the CelSeq2 protocol, mapping to bacterial sequences contributes up to 25% of the total reads-per-section. *Xenoturbella* is known to have a symbiotic relationship with the *Chlamydia* bacterium[220], but this contributes less than 20% of the overall bacterial sequences found. A much larger proportion of the bacterial sequences are instead found to map to various Proteobacteria, which have also been found in samples of *Xenoturbella* at great abundance[221]. A high proportion of bacterial reads in the sequencing data indicates that the inclusion of poly(A) selection is not sufficient to exclude bacterial RNA from analysis[222], and that more stringent alternative approaches (such as rRNA depletion) might be necessary to eliminate bacterial contamination. Nonetheless, the high number of total reads obtained from the CelSeq2 libraries means that

bacterial contamination has not hindered sequencing depth across the reads-of-interest from *Xenoturbella*.

8.3.4.2 Transcriptional signatures associated with anterior, gut, and posterior domains in Xenoturbella bocki

Data from all 90 sections were compared using PCA in order to search for any structure in common transcriptional signatures across the *Xenoturbella* AP axis (Figure 8.8). Unlike the *Xenoturbella* RNA-Seq data that was processed using exponential amplification in the SmartSeq2 protocol, libraries prepared using linear amplification in CelSeq2 do show a difference in their structure. As discussed, library 4 comprises problematic data, and this is shown in the distribution of sections representing posterior regions 1 (sections 61-70) and 2 (71-80) (post1 and post2, respectively). Post1 has a broad variation in composition (shown by the long blue oval in Figure 8.8), and Post2 is clearly very divergent from the rest of the data set, represented as the pink outlying cluster and suggesting a contamination problem.

The distribution of the remaining sections representing the anterior sections 1-30 (ant1, ant2, ant3); the middle 31-60 sections of *Xenoturbella* (mid1, mid2, mid3); and the final posterior sections 81-90 (post3) can be interpreted in line with what is known regarding *Xenoturbella* morphology. The distribution of sections in ant1 and ant2 (red and orange respectively) overlap with one another, which would be expected given that they come from the same region of the animal. The mid sections of the animal, visualised in three shades of green, cluster together very closely. From morphological analysis (see 6.1.2), we know that the central region of *Xenoturbella* is predominantly gut lumen and gastrodermis. Consequently, transcriptional signatures from the 30 sections representing the mid-section of the animal are likely to be commonly upregulated for the same digestive or gland-related genes (see section 7.2.3) and proportionally depleted for genes associated with epidermal, muscle or neuron cell types. In the PCA, sections

assigned to ant3 overlap with some of the sections found in the problematic post1 cluster. Although the anomalous sequencing results from library 4 mean that interpretation of this is less confident, similarity in gene expression between sections 21-30 and 61-70 would be likely, given that they represent tissue sections taken from either side of the central gut. Interestingly, sections representing post3 overlap in identity with those found in ant2, but not those from ant1. *In situ* hybridisation against neural genes (*XbElav*, see 6.2.1; and a *Xenoturbella*-specific neural gene, see 7.2.6) show specific upregulation at the very anterior of the animal. It is therefore possible that enrichment for neural genes makes the transcriptional signature of the first 10 sections more different from that of sections 11-20 and the final 10 sections.

8.3.4.3 Expression of Hox genes and AP neural markers

Heat maps generated for Hox genes and for genes associated with AP neural patterning in chordates[183] (*Elav*, *Musahashi*, *Sox1/2/3, Six3*, *Bf1*, *Dlx*, *nk2.1*, *Pax6*, *Tll*, *barH*, *Otx*, *Lim1/5*, *Gbx*, *Emx*, *Dbx*, *Vax*, *Rx*, *Irx* and *Engrailed*) did not show any patterns in differential expression across the AP axis in the CelSeq2-generated libraries, and many genes had 0 expression across all sections (Figure 8.9). The 0 coverage seen for some genes can likely be attributed to 'missing' hits to the current *Xenoturbella* gene models, as outlined in section 8.3.4.4, but it is also possible that the lack of any clear signal in data for Hox genes and the AP neural marker genes[183] could be attributed to the fact that an adult, and not embryonic, *Xenoturbella* animal was used for Tomoseq. The Hox and ParaHox group of genes are well studied and known to be involved in the regionalisation of the AP axis during development across the Bilateria[218]. However, this stereotypical AP expression of Hox genes is closely correlated with embryonic development. Evidence suggests that the expression of some of these genes persists into the adult in specific taxa[223], but the defined AP expression domains are most confidently associated with the embryonic stage.

Furthermore, putative orthologues of just four Hox genes (*Hox1*, *Hox4*, *Hox7* and *Hox11*) have been identified in *Xenoturbella* transcriptomic data. In the Acoelomorpha, three Hox genes have been identified, representing one each from an anterior, central, and posterior class[224]. Further investigation into the conservation of Hox and ParaHox genes in the Xenacoelomorpha is necessary to better understand their expression and function, but searching for specific domains of expression across the AP axis in the adult animal is perhaps not the most meaningful approach to uncover patterns of transcriptional signatures in the *Xenoturbella* CelSeq2 data.

8.3.4.4 Cell-type specific gene expression and Tomoseq data

Heat maps for genes associated with three specific cell types in *Xenoturbella* (neural cells, gland cells and muscle cells, see section 7.2.4) showed patterns of up-regulation across the AP axis that are inconsistent with results from *in situ* hybridisation. Whilst some neural cell genes and gland cell genes have patchy upregulation in individual sections across the data, these are not enhanced in the anterior or the midsection of the animal, as would be expected given the location of *Xenoturbella* neural and gland cells, respectively (Figure 8.10). Muscle-related genes (including myosin light and heavy chain) have a more ubiquitous expression across the AP heat map, which might be expected given the presence of a defined muscle layer underlying the epidermis of the animal in all axes (Figure 8.11, also see 6.1.2).

The heat map generated for the five ultrafiltratory genes (*XbNeph1*, *XbNephrin*, *XbPodocin-like*, *XbCD2AP* and *XbZO-1*) shows common upregulation of two or more genes from *XbPodocin-like*, *XbCD2AP* and *XbZO-1* in successive sections across the AP axis, with a broader region of expression of *XbPodocin-like* and *XbZO-1* near to the posterior of the animal (Figure 8.12). *XbNeph1* was not expressed across the 90 sections, and *XbNephrin1* appears to be expressed in just five individual sections, with two of these overlapping with *XbPodocin-like* expression.

The absence of any expression of *XbNeph1* is in common with the 0 coverage found for a number of other genes whose expression has been confirmed in *Xenoturbella* using *in situ* hybridisation. These include Synaptotagmin 1-like, all troponin sequences, and various digestion-related genes. The absence of reads in the sequencing data for genes that are known to be expressed in the adult animal can possibly be attributed to the 3' preferential sequencing of libraries amplified using CelSeq2. Consequently, reads for these genes are likely to correspond to the 3' UTR region, which is not adequately covered by the current gene models. Improving the 3' annotation of genes in the current genome is therefore necessary to identify these unmapped transcripts. This might also help to uncover more meaningful differential expression for specific cell types compared to the gene maps I have generated in this analysis. In addition, although linear amplification of RNA is reported to introduce fewer amplification biases than PCR-based exponential methods, some amplification artefacts are still likely to occur, which could hinder data interpretation.

8.4 General conclusions

In this analysis, I describe use of the Tomoseq protocol across the AP axis of *Xenoturbella*, and the RNA-Seq data that were obtained using two different RNA amplification protocols. I also describe the establishment of a linear amplification protocol in RNA extracted from small tissue sections of *Xenoturbella*.

Compared to exponential amplification, linear amplification of RNA extracted from cryosectioned tissue uncovered different transcriptomic signatures in regions across the anteroposterior of the animal, which can be explained in context with *Xenoturbella* morphology.

Heat maps for genes that have enhanced expression in specific cell types (neural, gland, muscle and ultrafiltratory cells) did not show a distribution across the AP axis that might have been expected given results

from single cell sequencing libraries and in situ hybridisation (see sections 7.2.4 and 7.2.6). Although linear amplification of RNA is reported to introduce fewer amplification biases than PCR-based exponential methods, some amplification artefacts are still likely to occur. It is therefore possible that preferential amplification of some transcripts might hinder data interpretation[214-216]. However, preferential amplification would not account for the 0 coverage values found for genes with known expression in *Xenoturbella*. This can potentially be attributed to the 3' preferential sequencing of the RNA libraries, resulting in many transcripts not being assigned to their correct gene model. Despite using Augustus gene prediction[225] and RNA-Seq models it is clear that annotation of the *Xenoturbella* genome needs to be improved for accurate differential expression inference. Subsequent analyses of the Tomoseq data will look to improve the 3' UTR annotation, in the hope that this might uncover meaningful expression patterns. For specific genes-of-interest, models could be manually extended in a 3' direction, but for improving annotation across the genome, the Augustus UTR prediction tool could also be implemented.

It is important to note that the analysis carried out in this chapter represents sequencing data from just one animal. In order for these findings - or any subsequent findings - to be reliable, further replicate animals will need to be used for Tomoseq assays. To verify the PCA that appears to show common gene signatures from groups of sections along the AP axis, I will also prepare libraries from sections which are not adjacent to one another, with the objective of overcoming any artefactual batch effects hindering confident data analysis.

A primary aim of subsequent analyses is to identify species-specific genes that have the most differential expression across the AP axis, but more than one biological replicate will be necessary in order to validate any variable gene expression. Furthermore, given that many of the prominent cell types in *Xenoturbella* (including for example, epidermal and muscle) are expressed in comparable layers across all three body axes, cryosectioning animals along the dorsoventral and sagittal axes will also be necessary to

fulfill the objective of building a 3D model of gene expression across the adult animal.

9 Discussion

9.1 General overview of initial objectives

The aim of this thesis was to use novel molecular approaches to investigate the biology of members of the enigmatic Xenacoelomorpha - an intriguing group of simple marine worms for which confident phylogenetic assignment remains challenging. Primarily, I framed this investigation within the context of investigating genes whose protein products are implicated in ultrafiltration in diverse taxa across the Bilateria: *Neph1*, *Nephrin*, *Podocin/EB7/Mec2, CD2AP* and *ZO-1*. Although these genes appear to have a conserved expression and function in a number of structurally divergent nephridial systems, there is no consensus regarding the origin and homology of excretory systems in bilaterians. However, the presence of a conserved molecular architecture at the site of ultrafiltration in the vertebrate podocytes and *D. melanogaster* nephrocytes provides a degree of evidence for the homology of these cells[130]. Furthermore, the expression of orthologues of two of these genes – *Smed-NPHS1-6*(=*Nephrin*) and *Smed-NEPH3*(=*Neph1*) - in the protonephridial flame cells of a flatworm (*S. mediterranea*) indicates that the core molecular components of ultrafiltration are conserved between metanephridial and protonephridial systems.

By investigating the presence and absence of these ultrafiltratory-related genes in bilaterian, diploblastic, and non-metazoan taxa, and using different molecular approaches to understand their expression in *Xenoturbella* and the acoel *S. roscoffensis*, I hoped to discover whether such cells exist in Xenacoelomorpha, and to advance the understanding of their function in the Xenacoelomorpha, and their wider historic function in the Metazoa.

In order to visualise expression patterns for the genes and proteins-of-interest, establishing reliable *in situ* hybridisation and immunohistochemistry

protocols in adult *Xenoturbella* was a primary objective. Perhaps given the difficulties associated with animal collection, and their inability to be kept in culture, no molecular protocols had previously been used in *Xenoturbella*. Although a number of molecular protocols have been established in the acoel members *I. pulchra* and *H. miamia*, reliable *in situ* hybridisation in *S. roscoffensis* – and in adult animals in particular – was a protocol that I also hoped to troubleshoot.

In addition to more traditional molecular approaches, I aimed to use innovative new RNA-Seq technologies in *Xenoturbella*. *Xenoturbella* are known to have a simple body plan, and histological and electron microscopy studies have been used to describe their morphology. However, the lack of comprehensive molecular data for this species means that very little was previously known regarding cell type complexity or specialisation in this species. By using whole organism single cell sequencing, I hoped to uncover transcriptional diversity across the cell population of *Xenoturbella* in a high-throughput pipeline, and employ computational clustering of the data to identify putative meta-clusters of cells, grouped by a common transcriptional signature. Not only did I aim to use this approach to identify co-expressed markers of ultrafiltration, but also to look much more widely to identify putative cell types in the adult animal to better describe their morphology. Complementarily to this approach, I attempted to establish RNA tomography (Tomoseq), to make cDNA libraries from tissue sections taken across the AP axis of *Xenoturbella*, with the objective of uncovering differential gene expression with a degree of spatial context.

Given the debate regarding the placement of the Xenacoelomorpha in the Bilateria - either basally in the Bilateria or as sister group of the Ambulacraria in the deuterostomes - the first chapter of this thesis focused on sequencing three mitochondrial genomes from across the Acoela. From this I aimed to contribute to the comparatively sparse molecular data available for this class of animals, and also use protein-coding gene sequences and other features of mitochondrial genomes for phylogenetic inference.

The experiments and analyses carried out in this thesis have achieved some of the objectives outlined above. Whilst molecular protocols in *S. roscoffensis* proved difficult to establish and to obtain confident results from, the molecular work carried out in *Xenoturbella* has contributed much more than was previously known regarding transcriptional diversity and the spatial distribution of cells in this species. In particular, *in situ* hybridisation results hint at the discovery of an interesting cell type, for which RNA probes for *XbNeph1*, *XbNephrin* and *XbPodocin-like* all show expression. In the following sections I will outline the main findings from the previous chapters in more detail, and the implications these might have for our understanding of the Xenacoelomorpha. Finally, I discuss how novel molecular approaches might aid our study of non-model organisms in the future, and how investigating unannotated or orphan genes might contribute to our understanding of cell type diversity and the biology of understudied taxa.

9.2 Acoela mitochondrial genomes and phylogenetic inference

9.2.1 New molecular data for the Acoela

In Chapter 3 I described the mitochondrial genomes from three different species of acoel: *P. rubra*, *I. pulchra* and *A. ylvae.* A number of different lines of evidence have been used to infer the phylogenetic position of the Xenacoelomorpha, including microRNAs, transcriptomic data, ESTs, and mitochondrial protein-coding genes, amongst others. Nonetheless, molecular data for this group are relatively poor. No complete nuclear genomes are currently published for any member of the Xenacoelomorpha, and mitochondrial data for the Acoela are sparse, with just one complete mitochondrial genome from *S. roscoffensis* previously published[19]. Consequently, the mitochondrial genomes described in this thesis contribute to the wealth of molecular data available for this class.

9.2.2 Novel structure of Acoela mitochondrial genomes

All three mitochondrial genomes that were sequenced have novel gene orders, both in comparison to one another, and to published mitochondrial genomes from across the Metazoa. Such prevalent gene order rearrangement could be a result of the rapid rate of sequence evolution in the Acoela, but the lack of any conserved blocks of genes means that gene order is not phylogenetically informative for this class, at least with the current depth of sampling. Further mitochondrial genome data from other acoel members would aid in this comparison. Interestingly, it seems possible from my data that the mitochondrial genome for *I. pulchra* has a duplicated region, which would provide evidence for a genomic 'duplication and deletion' rearrangement of genes in the mitochondrial genome of this species.

9.2.3 Phylogenetic inference using mitochondrial protein-coding genes

Given the absence of complete nuclear genomes for any member of the Xenacoelomorpha, mitochondrial protein-coding genes have been used in several analyses of phylogenetic inference for this phylum[23,127]. However, the lack of molecular data available for the Acoela means that they are represented in these analyses by just one complete mitochondrial genome (*S. roscoffensis*)[19]; one partially published mitochondrial genome (*P. rubra*)[20]; and the *cox1* genes from the acoel species *N. fusca* and *C. longifissura*. As has been found in other phylogenetic analyses of this phylum, mitochondrial data variably place the Xenacoelomorpha within the deuterostomes[23,28,127], or basally in the Bilateria[19].

In my analysis, all mitochondrial protein-coding genes found in all three species (*atp8* is absent in *I. pulchra* and *A. ylvae*; *nad4l* is also absent in *I. pulchra*) were used in both maximum likelihood and Bayesian phylogenetic inference. Acoela are known to have a faster-than-average rate of sequence evolution[28], which leaves them particularly vulnerable to LBA[107], and this has been cited as a reason for the basal position that is sometimes

inferred for them[28]. Indeed, in this analysis, the Acoela were found to group with the long branched Urochordata at the base of the tree, outside of the main protostome and deuterostome node. Excluding the Urochordata from analysis resolved the Acoela with Xenoturbella, placing the Xenacoelomorpha as a branch within the deuterostomes. This demonstrates the LBA problem that can confound phylogenetic inference for the position of the Acoela, and provides a degree of support for the basal position of Xenacoelomorpha resulting from a systematic error (LBA). However, it is clear that further molecular data from this phylum are needed for future phylogenetic analyses. Increased sampling of the Acoela to include more slowly evolving representatives could also help to 'break' the long branch. A more profitable source of data will be transcriptomes and genomes.

9.3 Establishing *in situ* hybridisation and immunohistochemistry protocols in *S. roscoffensis* and *Xenoturbella*

In order to fulfill the objective of visualising expression domains of genes that are known to have an ultrafiltratory function, reliable *in situ* hybridisation and immunohistochemistry protocols were necessary for a representative from the Acoela and for *Xenoturbella*.

9.3.1 Protocol troubleshooting in *S. roscoffensis*

Despite extensive troubleshooting of *in situ* hybridisation in adult *S. roscoffensis*, no protocol gave reliable, reproducible results for the RNA probes targeting the genes-of-interest. A possible exception to this was achieved using sectioned adult animals; taking this further will require experimental repeats with different probes to confirm. Whilst juvenile animals gave slightly more consistent and reproducible results, it appears that *S. roscoffensis* do not represent the most amenable species for *in situ* hybridisation protocols.

After establishing an optimal detergent concentration to permeabilise the epidermis of adult and juvenile *S. roscoffensis*, immunohistochemistry appeared to give more consistent signal than any expression patterns seen using *in situ* hybridisation. The likelihood of cross-reactivity with related proteins means that the specificity of signal from the antibodies used must be interpreted with caution, but this is not a fault of the protocol. Although some autofluorescence from the symbiont *T. convolutae* was evident in Z-stacks taken from the middle of the dorsoventral axis of the animal, positive signal from antibodies was consistently stronger than any endogenous autofluorescence.

Other acoels have been established in culture (*I. pulchra* and *H. miamia*), for which *in situ* hybridisation has been used with a degree of success. Nonetheless, maintaining these animals in culture can be problematic owing to frequent ciliate contamination problems and population crashes. As *S. roscoffensis* can be collected in abundant quantities, it was hoped that this species could present an alternative acoel to use in molecular protocols, without the problems associated with maintaining animal culture. Given the problems in implementing *in situ* hybridisation for the genes-of-interest, future investigations into gene expression in the Acoela might benefit from instead focusing on *I. pulchra* and *H. miamia*.

9.3.2 Molecular protocol establishment in *Xenoturbella*

As no molecular protocols have previously been used in *Xenoturbella,* establishing *in situ* hybridisation and immunohistochemistry protocols in this species was a primary objective of this thesis. As demonstrated by an initial control probe for *XbElav*, and subsequent verification as part of the single cell sequencing analysis, *in situ* hybridisation in *Xenoturbella* consistently gives expression patterns that are restricted to plausible and repeatable domains of expression for each gene-of-interest. *In situ* hybridisation is an enormously useful technique to have established in this species, and has

proved to be particularly valuable for providing a spatial context for putative cell type meta-clusters identified from the single cell sequencing pipeline.

The highly specific expression patterns observed for all of these probes means that I am also confident in the reliability of expression domains found for ultrafiltratory-genes, where there was no *a priori* knowledge about the cells in which they might be expressed.

9.4 Ultrafiltratory related genes in the Xenacoelomorpha

9.4.1 Presence and absence of ultrafiltratory-related genes in the Eukaryota

As outlined, the primary objective of my thesis was to identify the presence and expression of genes that are commonly associated with ultrafiltration across different bilaterian nephridial systems in members of the Xenacoelomorpha. My initial genome and transcriptome mining for these genes included publicly available data from across the Metazoa, and transcriptomes from *Xenoturbella* and from three acoels. The findings of these analyses largely support what has been previously published regarding the origin and distribution of these genes. It is evident that Neph1 and Nephrin, which form the core structure of the ultrafiltratory barrier in the vertebrate podocyte and *D. melanogaster* nephrocyte, are bilaterian novelties. No orthologues of *Neph1* or *Nephrin* were found in any non-bilaterian taxa, but were identified in all bilaterians included for analysis – including the Xenacoelomorpha. The distribution of three other genes with an ultrafiltratory role is not bilaterian-specific. Podocin is known to be a vertebrate-specific protein, but orthologues (EB7/Mec-2) were identified in the Xenacoelomorpha and throughout the Metazoa. Phylogenetic inference rejected the presence of *Podocin* orthologues outside of the Metazoa (in *C. owczarzaki* and *D. discoideum*). Similarly, *CD2AP* orthologues were identified in all metazoan taxa, but sequences with unclear orthology were

found for the Amoebozoa and Filasterea. Lastly, *ZO-1* orthologues were identified across the Eukaryota, with the exception of the Amoebozoa.

The presence of these genes across the Eukaryota indicates that *Podocin* (and respective orthologues), *CD2AP* and *ZO-1* did not have an ancestral ultrafiltratory function. Indeed all three genes are implicated in a number of roles across different cell types, and they are found in metazoan animals that are known to lack any ultrafiltratory or excretory system. More interesting is the identification of *Neph1* and *Nephrin* orthologues in members of the Xenacoelomorpha. Nephridial systems are regarded as a bilaterian novelty, and the co-expression of Neph1 and Nephrin proteins at the site of ultrafiltration appears to be well conserved across different bilaterian groups. Neph1 and Nephrin orthologues are found throughout the Bilateria - including taxa such as the tunicates and nematodes, which are thought to lack any ultrafiltratory or excretory system that can be defined as nephridia. Although both proteins are also thought to have roles in the patterning and function of the nervous system in vertebrates, it is their co-expression at the ultrafiltratory diaphragm that has been the focus of the majority of studies. Whether these proteins had an ancestral ultrafiltratory function or whether this role has been co-opted in separate lineages has not been extensively investigated. Although *C. elegans* orthologues of *Neph1*(=*SYG-1*) and *Nephrin*(=*SYG-2*) are reported to function at synapses[154], no analysis of the function or expression of Neph1 and Nephrin has been carried out in the tunicates.

9.4.2 Visualising ultrafiltratory genes in the Xenacoelomorpha

Difficulties with protocol establishment in *S. roscoffensis* means that expression patterns for *SrNeph1*, *SrNephrin* and *SrPodocin-like* are difficult to interpret reliably. As juvenile animals seemed to produce more consistent and reproducible results, the expression patterns seen for these genes are perhaps more reliable than those obtained in the adult. Nonetheless, expression of the three genes in the juvenile animal appears broader than might be expected for an ultrafiltratory function. *SrNeph1* appears to localise

more specifically to the parenchymal region flanking the gut, whilst *SrNephrin* and *SrPodocin-like* appear to be expressed throughout the gut region. Comparable domains of expression of ultrafiltratory genes have also been reported in the acoel *I. pulchra* (Andrikou *et al.* preprint)[158]. Nonetheless, although a location close to the gut or in the digestive system would not be unlikely for ultrafiltratory-related cells, the broad expression of these genes does not provide significant evidence for the presence of cells with an ultrafiltratory capacity.

Consistent results were achieved in *S. roscoffensis* using immunohistochemistry with species-specific polyclonal antibodies. However, the possibility of non-specific signal from conserved domains found in related proteins means that the putative nervous system signal seen for anti-SrNeph1 and anti-SrNephrin could be from other CAM-related proteins that are known to function in neural cells. Consequently, this signal is not particularly informative for informing the historic function of Neph1 and Nephrin proteins.

Owing to the mixed successes of gene visualisation approaches in *S. roscoffensis*, a functional analysis of these genes could be more informative in providing conclusive evidence for their role. RNAi in the flatworm *S. mediterranea* for transcription factors that are necessary for the patterning and maintenance of protonephridia resulted in visible bloating in the animal as a result of impaired excretion. As RNAi is already established in the acoels *I. pulchra* and *H. miamia*, targeting ultrafiltratory-related genes might prove an informative approach for understanding the role of these genes in the acoels.

9.4.3 Ultrafiltration and excretion in *Xenoturbella*

Unlike *S. roscoffensis*, implementing *in situ* hybridisation for the same three ultrafiltratory-related genes in *Xenoturbella* gave results that hint at a previously unknown cell type. All three genes (*XbNeph1*, *XbNephrin* and *XbPodocin-like*) appear to be expressed in cells located at the posterior

region of the animal, settling on the ECM that overlies the basal portion of the gastrodermis. The cells appear to be restricted to specific regions within the dorsoventral axis of the animal, although this varied between sections taken from two individuals. The ECM functions as a primary filtratory barrier in all nephridial systems, and so the location of these cells, overlying the ECM, lends a degree of support to a potential filtratory-related function. However, it will be necessary to verify the co-expression of these genes within the same cell by double *in situ* hybridisation. Immunogold labelling with species-specific antibodies may also aid in visualising signal – a technique that has been used in *D. melanogaster* nephrocytes for Sns (=Nephrin) and Duf (=Neph1)[130].

An interesting finding from these *in situ* hybridisation results was the difference in the number of cells and their distribution between sections and between individuals. Sections taken from across the dorsoventral axis of two individuals were used for each gene to investigate domains of expression. However, cells expressing these genes were only identified in a relatively narrow section of the dorsoventral axis for one of the individuals, covering a total thickness of ~140μM. This suggests that the cells-of-interest are only found in a specific region of this individual. More striking was the difference in the number of cells that were found between individuals. In the first set of *in situ* hybridisation experiments, cells that were found to express *XbNeph1*, *XbNephrin* and *XbPodocin-like* were distributed across the posterior region of the animal in great numbers. Conversely, in sections taken from a second individual, cells were far fewer, totalling just two or three cells per section. It is possible that these cells might be found in a denser concentration elsewhere in the dorsoventral axis, as outlined, but the difference in cell density between individuals could prove interesting to investigate further. Although members of the Urochordata are regarded to lack any ultrafiltratory or excretory systems that would be defined as nephridia, they do sequester nitrogenous waste and other metabolites in specialised vesicular cells[61]. It appears that there has not been extensive investigation into the structure of these cells, but they are thought to accumulate over the lifetime of the organism. Consequently, the number of vesicular cells to accumulate waste

would presumably increase in number over the lifespan of an organism. Although we have no reliable measure of the age of *Xenoturbella* individuals collected from the seabed, comparing cell number and distribution in sections taken from more individuals of different sizes as part of a double *in situ* hybridisation assay might provide support for the differences observed in cell number between individuals.

9.4.4 Re-interpreting the Nephrozoa

The discovery of these interesting cell types in *Xenoturbella* undoubtedly requires further investigation to uncover any ultrafiltratory-related capacity. Nevertheless, the expression of all three genes in discrete cells settling on the ECM could have obvious implications for the supposed 'Nephrozoa' clade that excludes Xenacoelomorpha.

My experiments in the Acoela do not demonstrate an ultrafiltratory-related function of cells expressing *SrNeph1* and *SrNephrin*. As has been described, however, the Nematoda and Urochordata lack nephridial-like structures, but have orthologues of Neph1 and Nephrin – and the absence of nephridia in these phyla have not prevented them from being encompassed within the Nephrozoa. It could be hypothesised that Neph1 and Nephrin, and their orthologues in respective bilaterians, had an ancestral function in the patterning and maintenance of the ultrafiltratory barrier: a role that has subsequently been lost in lineages that lack defined nephridial systems, such as the nematodes, tunicates and acoels. Although this would not provide evidence for either a basal bilaterian or deuterostome position of the Xenacoelomorpha, it would have implications for a revision of the Nephrozoa grouping, as has been outlined.

9.5 Novel RNA-Seq approaches in *Xenoturbella*

9.5.1 Success of single cell sequencing in *Xenoturbella*

Whole organism single cell sequencing is a novel approach for uncovering transcriptional diversity and putative cell types. For *Xenoturbella*, the preparation of single cell libraries from two different individuals resulted in the individual barcoding and sequencing of over 6000 cells, and the success of this protocol contributes much more than was previously known regarding the complexity of cell types in *Xenoturbella*. It is important to note that the single cell sequencing pipeline uses low coverage sequencing across a large population of cells. It is therefore likely that in clustering the data set, some cells might be artificially grouped into one population, or split into separate populations erroneously. In addition, the cells represented in the library will be in different states following enzymatic-induced dissociation - for example dying, dividing or bursting - and this will result in changes to their transcriptional signature. Nonetheless, investigating genes that are consistently enriched across a given meta-cluster can be used to assign a putative transcriptional identity to the cells therein.

Based on the transcriptional signatures identified in this analysis, meta-clusters of cells were assigned identities pertaining to neural, muscle, digestive gland, putative pigment, and epithelial and sensory epithelial cells. Six further meta-clusters of cells could not be assigned a confident cell-type identity based on their transcriptional signature, but were variably enriched for genes with expression and function in numerous cell types, including mitochondrial and ribosomal genes, and transferases. *In situ* hybridisation to verify the single cell sequencing data correlated successfully with the expected domains of expression. Neural, sensory, muscle and gut/digestive tissue-types are found throughout the Metazoa, but the cell-specific upregulation of different genes within these meta-clusters is indicative of additional hidden diversity within these cell types. A surprising finding from the single cell libraries is the large contribution of sensory epithelial cells to the dataset. These cells are commonly enriched for *Xenoturbella*-specific

genes. It has been assumed that the nervous system in *Xenoturbella* is a simple nerve net, lacking centralisation or diverse neural types. However, the cell diversity within the neural meta-cluster, and the common upregulation of genes found in nerve cells and putative sensory epithelial cells indicates more complexity than was previously thought. These patterns of expression also suggest that *Xenoturbella* has neural-like sensory cells that could have a species-specific function, but are found outside of the main basiepithelial nerve net.

9.5.2 Future analysis of the single cell data

The results described in this thesis represent only an initial analysis into the single cell sequencing libraries assembled from *Xenoturbella*. Having assigned identity to six meta-clusters represented in the data, I aim to re-cluster these populations further to uncover additional transcriptional diversity within these cell populations. In particular, the number of *Xenoturbella*-specific orphan genes with strong meta-cluster identity could prove interesting for investigation. Given the success of *in situ* hybridisation in identifying the expression of meta-cluster enriched genes in specific cells, targeting these orphan genes for *in situ* hybridisation would give a degree of spatial context for their expression and hopefully tell us more about the cell types identified based on unique transcriptional profiles.

In addition, further analysis into the enrichment of transcription factors across the dataset is a primary objective to uncover patterns of transcriptional regulation. Preliminary analysis of cell clusters associated with selected transcription factors uncovered a degree of meta-cluster specific enrichment - for example, the expression of *ISL1* in the neural meta-cluster. Further investigation into the degree of meta-cluster specific or cell specific transcription factor expression could provide insight into a hierarchy of transcriptional regulation within putative tissue types in *Xenoturbella*. For example, the gene repertoires of the neural, sensory epithelial and gland cell meta-clusters in particular suggest a diversity of cell types within these

broader groupings. Investigating the degree of meta-cluster specificity of transcription factors expressed in these cell-types could help uncover the regulatory mechanisms associated with this diversity. High transcription factor diversity might reflect additional transcriptional organisation in these meta-clusters, thus contributing to the cell type diversity observed in this analysis. Analysing the transcription factor complement of cells enriched for ribosomal and cell-cycle genes (currently in unassigned meta-clusters) could also help inform our understanding of putative neoblast differentiation in *Xenoturbella*.

9.5.3 Tomoseq and linear amplification in *Xenoturbella*

My initial objective with regards to spatial transcriptomics in *Xenoturbella* was to establish RNA tomography along the anteroposterior axis of the animal, in order to investigate possible differential gene expression. It was apparent from initial sequencing data that exponential amplification of RNA extracted from the tissue sections could be masking any variation in transcription along this body axis. Consequently, I aimed to establish a linear amplification protocol for RNA extracted from tissue slices of *Xenoturbella*, as part of the Tomoseq pipeline. Establishing a reliable RNA extraction technique from small tissue sections of *Xenoturbella* required a number of troubleshooting approaches. Having optimised this protocol, further protocol refinement was necessary to implement CelSeq2 using the clean RNA. Using spike-in RNA-Seq reads, I was able to verify the success of linear amplification of the RNA – a novel approach to have established in *Xenoturbella*, and a strategy that could also be applicable to other non-model organisms.

9.5.4 Differential gene expression across the anteroposterior axis of *Xenoturbella*

PCA of the initial, exponentially amplified RNA-Seq data did not find any variation between AP sections, and it appears that PCR-based RNA

amplification approaches could thus prove problematic for investigating differential gene expression. PCA of RNA-Seq data from the linear amplification protocol, however, showed variation across the AP axis of the animal, and this variation can be cautiously interpreted in line with what we know of *Xenoturbella* morphology.

Heat maps of expression for different genes-of-interest gave less conclusive findings. For Hox genes and other chordate AP patterning genes, variation across the AP axis did not vary as might be expected for a bilaterian organism. However, to uncover meaningful patterns of expression of these developmental genes, using the Tomoseq approach on an embryonic *Xenoturbella* would likely be the optimal strategy. As *Xenoturbella* cannot be maintained in culture, having access to an embryonic animal for this protocol is unlikely.

I next focused analysis on genes that I found to have specific domains of expression in the single cell sequencing pipeline. Patchy upregulation of expression across the AP axis was found for some neural and gland cell related genes, whilst some muscle related genes had more ubiquitous expression across all sections. However, for some genes whose expression has been verified by *in situ* hybridisation, RNA-Seq data from Tomoseq showed no expression, perhaps indicating that reads are not being mapped to their correct gene model as a result of preferential 3' UTR sequencing. For subsequent analysis of the Tomoseq data, improving gene model annotation at the 3' end of genes is a clear priority: for specific genes-of-interest, models could be manually extended in a 3' orientation, but for improving annotation across the genome, the Augustus UTR gene model prediction tool should also be implemented.

9.5.5 Future strategies for Tomoseq implementation and analysis

The establishment of the sectioning, linear amplification, and cDNA library preparation protocols contributes to the molecular approaches that are

now refined for *Xenoturbella*. As outlined, the future objective for Tomoseq RNA-Seq data analysis is to improve the 3' annotation of *Xenoturbella* gene models. This will hopefully contribute to more meaningful analysis of gene expression across the AP axis of the adult animal. Having established the Tomoseq pipeline using successful linear amplification, this protocol will also need repeating in more *Xenoturbella* individuals – along the AP axis to provide biological replicates but also along the dorsoventral and left-right body axes with the ultimate objective of building a 3D model of gene expression in *Xenoturbella*. The establishment of the Tomoseq and CelSeq2 protocol in *Xenoturbella* could also prove valuable for investigating domains of gene expression in other non-model organisms.

9.6 RNA-Seq approaches in evo-devo

The encompassing objective of this thesis was to apply molecular approaches to investigate the biology of members of the Xenacoelomorpha. Whilst initial molecular approaches (*in situ* hybridisation and immunohistochemistry) were used on a member of the Acoela (*S. roscoffensis*) and on *Xenoturbella*, subsequent projects focused on using novel RNA-Seq approaches exclusively in the enigmatic species *Xenoturbella*. Techniques such as single cell sequencing and Tomoseq are innovative and exciting, and can generate a huge amount of data very rapidly. From implementing whole organism single cell sequencing in *Xenoturbella*, I was able to identify gene expression patterns and putative cell type diversity far more rapidly than would have been possible by traditional gene visualisation techniques. *In situ* hybridisation remains a valuable approach for visualising gene expression – and is necessary to validate RNA-Seq data – but for non-model organisms such as *Xenoturbella*, employing these high-throughput, novel sequencing approaches will provide an approach for uncovering gene expression data very rapidly. In addition, it is clear from the *Xenoturbella* single cell libraries that the expression of *Xenoturbella*-specific orphan genes is often specific to certain cells within a broader cell-type cluster. As orphan genes are thought to function in lineage-

specific adaptations, single cell data could also help uncover novel cell types enriched for orphan gene expression in understudied taxa that would not be identifiable by the standard candidate gene approach. Nonetheless, it is clear from the data analysis outlined in this thesis that gene annotation is of paramount importance for the successful interpretation of RNA-Seq data. For many non-model taxa, genome sequencing and assembly, with reliable gene model annotation, would be necessary prior to the implementation of these techniques.

Although RNA-Seq technologies such as single cell sequencing and RNA tomography are currently in their infancy with regard to whole-organism application, they represent the wealth of molecular techniques that we can now use to compare and understand more about the diversity of the Metazoa. As outlined at the start of this thesis, evo-devo is a fundamentally comparative science: we cannot exclusively use model organisms to infer the expression and function of genes across taxa that are separated by millions of years of evolution. Using these new technologies could contribute a great deal more to what is currently known of understudied taxa – and with that, broaden our understanding of ancestral gene functions and cell type specifications.

References

1 Haeckel, E. *Generelle Morphologie der Organismen : allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*. (Berlin : G. Reimer, 1866., 1866).

2 Telford, M. J., Budd, G. E. & Philippe, H. Phylogenomic Insights into Animal Evolution. *Current Biology* **25**, R876-887, doi:10.1016/j.cub.2015.07.060 (2015).

3 Telford, M. J. & Copley, R. R. Zoology: War of the Worms. *Current Biology* **26**, R335-337, doi:10.1016/j.cub.2016.03.015 (2016).

4 Edgecombe, G. D. *et al.* Higher-level metazoan relationships: recent progress and remaining questions. *Organisms Diversity & Evolution* **11**, 151-172 (2011).

5 Bourlat, S. J., Nielsen, C., Economou, A. D. & Telford, M. J. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. *Molecular Phylogenetics and Evolution* **49**, 23-31, doi:S1055-7903(08)00362-X [pii]
10.1016/j.ympev.2008.07.008 (2008).

6 Dunn, C. W., Giribet, G., Edgecombe, G. D. & Hejnol, A. Animal Phylogeny and Its Evolutionary Implications. *Annual Review of Ecology, Evolution, and Systematics* **45**, 371-395, doi:10.1146/annurev-ecolsys-120213-091627 (2014).

7 Adoutte, A. *et al.* The new animal phylogeny: reliability and implications. *Proceedings of the National Academy of Sciences* **97**, 4453-4456, doi:97/9/4453 [pii] (2000).

8 Giribet, G. New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. *Organisms Diversity & Evolution* **16**, 419-426, doi:10.1007/s13127-015-0236-4 (2016).

9 Bourlat, S. J. & Hejnol, A. Acoels. *Current Biology* **19**, R279-280, doi:10.1016/j.cub.2009.02.045 (2009).

10 Achatz, J. G., Chiodin, M., Salvenmoser, W., Tyler, S. & Martinez, P. The Acoela: on their kind and kinships, especially with nemertodermatids and xenoturbellids (Bilateria incertae sedis). *Organisms, Diversity & Evolution* **13**, 267-286, doi:10.1007/s13127-012-0112-4 (2013).

11 Perea-Atienza, E. *et al.* Posterior regeneration in *Isodiametra pulchra* (Acoela, Acoelomorpha). *Frontiers in Zoology* **10**, 64, doi:10.1186/1742-9994-10-64 (2013).

12 Wijgerde, T. *et al.* Epizoic acoelomorph flatworms impair zooplankton feeding by the scleractinian coral *Galaxea fascicularis*. *Biology Open* (2012).

13 Ruiz-Trillo, I., Riutort, M., Littlewood, D. T., Herniou, E. A. & Baguña, J. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* **283**, 1919-1923 (1999).

14 Katayama, T., Yamamoto, M., Wada, H. & Satoh, N. Phylogenetic position of Acoel turbellarians inferred from partial 18S rDNA sequences. *Zoological Science* **10**, 529-536 (1993).

15 Philippe, H., Brinkmann, H., Martinez, P., Riutort, M. & Baguñà, J. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *PLoS One* **2**, e717, doi:10.1371/journal.pone.0000717 (2007).
16 de Rosa, R. *et al.* Hox genes in brachiopods and priapulids and protostome evolution. *Nature* **399**, 772-776, doi:10.1038/21631 (1999).
17 Cook, C. E., Jiménez, E., Akam, M. & Saló, E. The Hox gene complement of acoel flatworms, a basal bilaterian clade. *Evol Dev* **6**, 154-163, doi:10.1111/j.1525-142X.2004.04020.x (2004).
18 Telford, M. J., Lockyer, A. E., Cartwright-Finch, C. & Littlewood, D. T. Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 1077-1083, doi:10.1098/rspb.2003.2342 (2003).
19 Mwinyi, A. *et al.* The phylogenetic position of Acoela as revealed by the complete mitochondrial genome of *Symsagittifera roscoffensis*. *BMC Evolutionary Biology* **10**, 309, doi:10.1186/1471-2148-10-309 (2010).
20 Ruiz-Trillo, I., Riutort, M., Fourcade, H. M., Baguñà, J. & Boore, J. L. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Molecular Phylogenetics and Evolution* **33**, 321-332, doi:10.1016/j.ympev.2004.06.002 (2004).
21 Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci* **276**, 4261-4270, doi:10.1098/rspb.2009.0896 (2009).
22 Westblad, E. Xenoturbella bockin.g., n. sp. a peculiar, primitive Turbellarian type. *Arkiv för zoologi* **1** (1949).
23 Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. *Nature* **530**, 94-97, doi:10.1038/nature16545 http://www.nature.com/nature/journal/v530/n7588/abs/nature16545.html - supplementary-information (2016).
24 Noren, M. & Jondelius, U. Xenoturbella's molluscan relatives. *Nature* **390**, doi:10.1038/36242 (1997).
25 Bourlat, S. J., Nielsen, C., Lockyer, A. E., Littlewood, D. T. & Telford, M. J. Xenoturbella is a deuterostome that eats molluscs. *Nature* **424**, 925-928, doi:10.1038/nature01851 (2003).
26 Lundin, K. The epidermal ciliary rootlets of *Xenoturbella bocki* (Xenoturbellida) revisited: new support for a possible kinship with the Acoelomorpha (Platyhelminthes). *Zoologica Scripta* **27**, 8 (1998).
27 Raikova, O. I., Reuter, M., Jondelius, U. & Gustafsson, M. K. The brain of the Nemertodermatida (Platyhelminthes) as revealed by anti-5HT and anti-FMRFamide immunostainings. *Tissue Cell* **32**, 358-365, doi:10.1054/tice.2000.0121 (2000).
28 Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature* **470**, 255-258, doi:10.1038/nature09676 (2011).

29 Srivastava, M., Mazza-Curll, Kathleen L., van Wolfswinkel, Josien C. & Reddien, Peter W. Whole-Body Acoel Regeneration Is Controlled by Wnt and Bmp-Admp Signaling. *Current Biology* **24**, 1107-1113, doi:http://dx.doi.org/10.1016/j.cub.2014.03.042 (2014).
30 Cannon, J. T. *et al.* Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530**, 89-93, doi:10.1038/nature16520 http://www.nature.com/nature/journal/v530/n7588/abs/nature16520.html - supplementary-information (2016).
31 DeSalle, R. & Schierwater, B. Key transitions in animal evolution. *Integrative and Comparative Biology* **47**, 667-669, doi:10.1093/icb/icm042 (2007).
32 Rosslenbroich, B. The evolution of multicellularity in animals as a shift in biological autonomy. *Theory in Biosciences* **123**, 243-262, doi:10.1016/j.thbio.2004.10.002 (2005).
33 Wright, P. A. Nitrogen excretion: three end products, many physiological roles. *The Journal of Experimental Biology* **198**, 273-281 (1995).
34 Scimone, M. L., Srivastava, M., Bell, G. W. & Reddien, P. W. A regulatory program for excretory system regeneration in planarians. *Development* **138**, 4387-4398, doi:10.1242/dev.068098 (2011).
35 Raciti, D. *et al.* Organization of the pronephric kidney revealed by large-scale gene expression mapping. *Genome Biology* **9**, R84-R84, doi:10.1186/gb-2008-9-5-r84 (2008).
36 Hediger, M. A. *et al.* The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. *Pflügers Archiv* **447**, 465-468, doi:10.1007/s00424-003-1192-y (2004).
37 Thi-Kim Vu, H. *et al.* Stem cells and fluid flow drive cyst formation in an invertebrate excretory organ. *Elife* **4**, doi:10.7554/eLife.07405 (2015).
38 Agarwal, S. K. & Gupta, A. Aquaporins: The renal water channels. *Indian Journal of Nephrology* **18**, 95-100, doi:10.4103/0971-4065.43687 (2008).
39 McKanna, J. A. Fine structure of the protonephridial system in Planaria. I. Flame cells. *Z Zellforsch Mikrosk Anat* **92**, 509-523 (1968).
40 Ruppert, E. E. & Smith, P. R. The functional organization of filtration nephridia. *Biological Reviews* **63**, 231-258, doi:10.1111/j.1469-185X.1988.tb00631.x (1988).
41 Bartolomaeus, T. & Ax, P. Protonephridia and Metanephridia - their relation within the Bilateria. *Journal of Zoological Systematics and Evolutionary Research* **30**, 21-45, doi:10.1111/j.1439-0469.1992.tb00388.x (1992).
42 Reece, J. B. C., N. A. . *Biology*. 7 edn, (Benjamin Cummings/Pearson, 2004).
43 Wilson, R. A. & Webster, L. A. Protonephridia. *Biological Reviews* **49**, 127-160, doi:10.1111/j.1469-185X.1974.tb01572.x (1974).
44 Hertel, L. A. Excretion and Osmoregulation in the Flatworms. *Transactions of the American Microscopical Society* **112**, 10-17, doi:10.2307/3226778 (1993).

45 Mattern, C. F. T. & Daniel, W. A. The Flame Cell of Rotifer: Electron Microscope Observations of Supporting Rootlet Structures. *The Journal of Cell Biology* **29**, 552-554 (1966).
46 Norrevang, A. Fine Structure of the Solenocyte Tree in Priapulus caudatus Lamarck. *Nature* **198**, 700-701 (1963).
47 Kieneke, A., Arbizu, P. M. & Ahlrichs, W. H. Ultrastructure of the protonephridial system in *Neodasys chaetonotoideus* (Gastrotricha: Chaetonotida) and in the ground pattern of Gastrotricha. *Journal of Morphology* **268**, 602-613, doi:10.1002/jmor.10536 (2007).
48 Brandenburg, J. Die reusenzelle (cyrtocyte) des Dinophilus (archiannelida). *Zeitschrift für Morphologie der Tiere* **68**, 83-92, doi:10.1007/BF00277424 (1970).
49 Patrakka, J. *et al.* Congenital nephrotic syndrome (NPHS1): Features resulting from different mutations in Finnish patients. *Kidney International* **58**, 972-980 (2000).
50 Hansen, U. Electron microscopic study of possible sites of ultrafiltration in *Lumbricus terrestris* (Annelida, Oligochaeta). *Tissue and Cell* **28**, 195-203, doi:http://dx.doi.org/10.1016/S0040-8166(96)80007-3 (1996).
51 Boer, H. H. & Sminia, T. Sieve structure of slit diaphragms of podocytes and pore cells of gastropod molluscs. *Cell and Tissue Research* **170**, 221-229, doi:10.1007/BF00224300 (1976).
52 Storch, V. & Herrmann, K. Podocytes in the blood vessel linings of *Phoronis muelleri* (Phoronida, Tentaculata). *Cell and Tissue Research* **190**, 553-556, doi:10.1007/bf00219564 (1978).
53 Balser, E. J. & Ruppert, E. E. Structure, Ultrastructure, and Function of the Preoral Heart-Kidney in Saccoglossus kowalevskii (Hemichordata, Enteropneusta) Including New Data on the Stomochord. *Acta Zoologica* **71**, 235-249, doi:10.1111/j.1463-6395.1990.tb01082.x (1990).
54 Welsch, U. & Rehkämper, G. Podocytes in the axial organ of echinoderms. *Journal of Zoology* **213**, 45-50, doi:10.1111/j.1469-7998.1987.tb03675.x (1987).
55 Ruppert, E. E. Morphology of Hatschek's Nephridiudm in Larval and Juvenile Stages of *Branchiostoma virginiae* (Cephalochordata). *Israel Journal of Zoology* **42**, S161-S182, doi:10.1080/00212210.1996.10688879 (1996).
56 Stach, T. & Eisler, K. The Ontogeny of the Nephridial System of the Larval Amphioxus (*Branchiostoma lanceolatum*). *Acta Zoologica* **79**, 113-118, doi:10.1111/j.1463-6395.1998.tb01150.x (1998).
57 Ruppert, E. E. Evolutionary Origin of the Vertebrate Nephron. *American Zoologist* **34**, 542-553, doi:10.2307/3883863 (1994).
58 Schmidt-Rhaesa, A. *The Evolution of Organ Systems*. (Oxford University Press, 2007).
59 Turpenniemi, T. A. & Hyvärinen, H. Structure and Role of the Renette Cell and Caudal Glands in the Nematode Sphaerolaimus gracilis (Monhysterida). *Journal of Nematology* **28**, 318-327 (1996).
60 Sundaram, M. V. & Buechner, M. The Caenorhabditis elegans Excretory System: A Model for Tubulogenesis, Cell Fate Specification, and Plasticity. *Genetics* **203**, 35-63 (2016).

61 Lacalli, T. C. Tunicate tails, stolons, and the origin of the vertebrate trunk. *Biological Reviews* **74**, 177-198 (1999).
62 Ruppert, E. E. Structure, Ultrastructure and Function of the Neural Gland Complex of *Ascidia interrupta* (Chordata, Ascidiacea): Clarification of Hypotheses Regarding the Evolution of the Vertebrate Anterior Pituitary. *Acta Zoologica* **71**, 135-149, doi:10.1111/j.1463-6395.1990.tb01189.x (1990).
63 Dunagan, T. T. & Miller, D. M. A Review of Protonephridial Excretory Systems in Acanthocephala. *The Journal of Parasitology* **72**, 621-632, doi:10.2307/3281449 (1986).
64 Funch, P. The chordoid larva of *Symbion pandora* (Cycliophora) is a modified trochophore. *Journal of Morphology* **230**, 231-263, doi:10.1002/(SICI)1097-4687(199612)230:3<231::AID-JMOR1>3.0.CO;2-H (1996).
65 Bartolomaeus, T. Ultrastructure and relationship between protonephridia and metanephridia in *Phoronis muelleri* (Phoronida). *Zoomorphology* **109**, 113-122, doi:10.1007/BF00312317 (1989).
66 Bartolomaeus, T. & Quast, B. in *Morphology, Molecules, Evolution and Phylogeny in Polychaeta and Related Taxa* Vol. 179 *Developments in Hydrobiology* (eds T. Bartolomaeus & G. Purschke) Ch. 9, 139-165 (Springer Netherlands, 2005).
67 Baeumler, N., Haszprunar, G. & Ruthensteiner, B. Development of the excretory system in a polyplacophoran mollusc: stages in metanephridial system development. *Frontiers in Zoology* **9**, 1-17, doi:10.1186/1742-9994-9-23 (2012).
68 Goodrich, E. S. The study of nephridia and genital ducts since 1895. *Quarterly Journal of Microscopical Science* **86**, 113-301 (1945).
69 Goodrich, E. S. The Early Development of the Nephridia in Amphioxus: Introduction and part I, Hattschek's nephridium. *Quarterly Journal of Microscopical Science* **76**, 499-510 (1934).
70 Achatz, J. G. & Martinez, P. The nervous system of Isodiametra pulchra (Acoela) with a discussion on the neuroanatomy of the Xenacoelomorpha and its evolutionary implications. *Frontiers in Zoology* **9**, 27, doi:10.1186/1742-9994-9-27 (2012).
71 De Mulder, K. *et al.* Characterization of the stem cell system of the acoel Isodiametra pulchra. *BMC Developmental Biology* **9**, 1-17, doi:10.1186/1471-213x-9-69 (2009).
72 Egger, B. *et al.* To be or not to be a flatworm: the acoel controversy. *Plos One* **4** (2009).
73 Carroll, S. B. Chance and necessity: the evolution of morphological complexity and diversity. *Nature* **409**, 1102-1109 (2001).
74 Junker, Jan P. *et al.* Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell* **159**, 662-675, doi:http://dx.doi.org/10.1016/j.cell.2014.09.038 (2014).
75 Guillard, R. R. L. & Ryther, J. H. STUDIES OF MARINE PLANKTONIC DIATOMS: I. CYCLOTELLA NANA HUSTEDT, AND DETONULA CONFERVACEA (CLEVE) GRAN. *Canadian Journal of Microbiology* **8**, 229-239, doi:10.1139/m62-029 (1962).
76 Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes.

Bioinformatics **29**, 2790-2791, doi:10.1093/bioinformatics/btt468 (2013).
77 Bernt, M. *et al.* MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution* **69**, 313-319, doi:10.1016/j.ympev.2012.08.023 (2013).
78 Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38**, W7-W13, doi:10.1093/nar/gkq291 (2010).
79 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539-539, doi:10.1038/msb.2011.75 (2011).
80 Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973, doi:10.1093/bioinformatics/btp348 (2009).
81 Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, doi:10.1093/molbev/msj030 (2006).
82 Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, doi:10.1093/bioinformatics/btu033 (2014).
83 Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* **21**, doi:10.1093/molbev/msh112 (2004).
84 Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649, doi:10.1093/bioinformatics/bts199 (2012).
85 Zhang, Z. *et al.* ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* **419**, 779-781, doi:http://dx.doi.org/10.1016/j.bbrc.2012.02.101 (2012).
86 Zhang, Z. *et al.* KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. *Genomics, Proteomics & Bioinformatics* **4**, 259-263, doi:http://dx.doi.org/10.1016/S1672-0229(07)60007-2 (2006).
87 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, doi:10.1093/molbev/mst010 (2013).
88 Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**, 776 (2014).
89 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495-502, doi:10.1038/nbt.3192
http://www.nature.com/nbt/journal/v33/n5/abs/nbt.3192.html - supplementary-information (2015).

90 Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, doi:10.1038/nmeth.2639 (2013).
91 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525-527, doi:10.1038/nbt.3519
http://www.nature.com/nbt/journal/v34/n5/abs/nbt.3519.html - supplementary-information (2016).
92 Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods* **14**, 687-690, doi:10.1038/nmeth.4324
http://www.nature.com/nmeth/journal/v14/n7/abs/nmeth.4324.html - supplementary-information (2017).
93 Robertson, H. E., Lapraz, F., Egger, B., Telford, M. J. & Schiffer, P. H. The mitochondrial genomes of the acoelomorph worms *Paratomella rubra*, *Isodiametra pulchra* and *Archaphanostoma ylvae*. *Scientific Reports* **7**, 1847, doi:10.1038/s41598-017-01608-4 (2017).
94 Robertson, H. E., Lapraz, F., Rhodes, A. C. & Telford, M. J. The Complete Mitochondrial Genome of the Geophilomorph Centipede *Strigamia maritima*. *PLOS ONE* **10**, e0121369, doi:10.1371/journal.pone.0121369 (2015).
95 Chipman, A. D. *et al.* The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede *Strigamia maritima*. *PLoS Biology* **12**, e1002005, doi:10.1371/journal.pbio.1002005 (2014).
96 Bourlat, S. J., Rota-Stabelli, O., Lanfear, R. & Telford, M. J. The mitochondrial genome structure of *Xenoturbella bocki* (phylum Xenoturbellida) is ancestral within the deuterostomes. *BMC Evol Biol* **9**, doi:10.1186/1471-2148-9-107 (2009).
97 Telford, M. J., Herniou, E. A., Russell, R. B. & Littlewood, D. T. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proceedings of the National Academy of Sciences* **97**, 11359-11364, doi:10.1073/pnas.97.21.11359 (2000).
98 Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. & Reyes, A. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* **238**, 195-209 (1999).
99 Boore, J. L. & Brown, W. M. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Current Opinion in Genetics & Development* **8**, 668-674 (1998).
100 Moritz, C. & Brown, W. M. Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. *Proceedings of the National Academy of Sciences* **84**, 7183-7187 (1987).
101 Boore, J. L., Lavrov, D. V. & Brown, W. M. Gene translocation links insects and crustaceans. *Nature* **392**, 667-668, doi:10.1038/33577 (1998).
102 Gibson, A., Gowri-Shankar, V., Higgs, P. G. & Rattray, M. A comprehensive analysis of mammalian mitochondrial genome base

composition and improved phylogenetic methods. *Molecular Biology and Evolution* **22**, 251-264, doi:10.1093/molbev/msi012 (2005).
103 Reyes, A., Gissi, C., Pesole, G. & Saccone, C. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology and Evolution* **15**, 957-966 (1998).
104 Perna, N. T. & Kocher, T. D. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *Journal of Molecular Evolution* **41**, 353-358 (1995).
105 Foster, P. G. & Hickey, D. A. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution* **48**, 284-290 (1999).
106 Hassanin, A. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. *Molecular Phylogenetics and Evolution* **38**, 100-116, doi:10.1016/j.ympev.2005.09.012 (2006).
107 Felsenstein, J. Cases in which parsimony or compatability methods will be positively misleading. *Systematic Zoology* **27**, 401-410 (1978).
108 Carranza, S., Baguñà, J. & Riutort, M. Are the Platyhelminthes a monophyletic primitive group? An assessment using 18S rDNA sequences. *Molecular Biology and Evolution* **14**, 485-497 (1997).
109 Crezee, M. *Paratomella rubra*, Rieger and Ott, an amphiatlantic acoel turbellarian. *Cahiers De Biologie Marine* **19**, 1-9 (1978).
110 Rieger, R. & Ott, J. Gezeitenbedingte Wanderungen von Turbellarien und Nematoden eines nordadriatischen Sandstrandes. *Vie Milieu (Suppl.)* **22**, 425-447 (1971).
111 Kånneby, T., Bernvi, D. C. & Jondelius, U. Distribution, delimitation and description of species of *Archaphanostoma* (Acoela). *Zoologica Scripta* **44**, 218-231, doi:10.1111/zsc.12092 (2015).
112 Bernt, M. *et al.* CREx: inferring genomic rearrangements based on common intervals. *Bioinformatics* **23**, 2957-2958, doi:10.1093/bioinformatics/btm468 (2007).
113 Hebert, P. D. N., Ratnasingham, S. & de Waard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, S96-S99, doi:10.1098/rsbl.2003.0025 (2003).
114 Sakai, M. & Sakaizumi, M. The complete mitochondrial genome of *Dugesia japonica* (Platyhelminthes; order Tricladida). *Zoological Science* **29**, 672-680, doi:10.2108/zsj.29.672 (2012).
115 Moritz, C., Dowling, T. E. & Brown, W. M. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics* **18**, doi:10.1146/annurev.es.18.110187.001413 (1987).
116 Braband, A., Podsiadlowski, L., Cameron, S. L., Daniels, S. & Mayer, G. Extensive duplication events account for multiple control regions and pseudo-genes in the mitochondrial genome of the velvet worm *Metaperipatus inae* (Onychophora, Peripatopsidae). *Molecular Phylogenetics and Evolution* **57**, 293-300, doi:http://dx.doi.org/10.1016/j.ympev.2010.05.012 (2010).

117 Hyman, B. C., Beck, J. L. & Weiss, K. C. Sequence amplification and gene rearrangement in parasitic nematode mitochondrial DNA. *Genetics* **120**, 707-712 (1988).
118 Zhou, X., Lin, Q., Fang, W. & Chen, X. The complete mitochondrial genomes of sixteen ardeid birds revealing the evolutionary process of the gene rearrangements. *BMC Genomics* **15**, 573, doi:10.1186/1471-2164-15-573 (2014).
119 Hyman, B. C., Lewis, S. C., Tang, S. & Wu, Z. Rampant gene rearrangement and haplotype hypervariation among nematode mitochondrial genomes. *Genetica* **139**, 611-615, doi:10.1007/s10709-010-9531-3 (2011).
120 Bermudez-Santana, C. *et al.* Genomic organization of eukaryotic tRNAs. *BMC Genomics* **11**, 1-14, doi:10.1186/1471-2164-11-270 (2010).
121 Fonseca, V. G. *et al.* Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications* **1**, 98, doi:10.1038/ncomms1095
http://www.nature.com/articles/ncomms1095 - supplementary-information (2010).
122 Schiffer, P. H. & Herbig, H.-G. Endorsing Darwin: global biogeography of the epipelagic goose barnacles Lepas spp. (Cirripedia, Lepadomorpha) proves cryptic speciation. *Zoological Journal of the Linnean Society* **177**, 507-525, doi:10.1111/zoj.12373 (2016).
123 Morrison, D. A. How and where to look for tRNAs in Metazoan mitochondrial genomes, and what you might find when you get there. *arXiv.org* (2010).
124 He, Y., Jones, J., Armstrong, M., Lamberti, F. & Moens, M. The Mitochondrial Genome of *Xiphinema americanum sensu stricto* (Nematoda: Enoplea): Considerable Economization in the Length and Structural Features of Encoded Genes. *Journal of Molecular Evolution* **61**, 819-833, doi:10.1007/s00239-005-0102-7 (2005).
125 Minxiao, W., Song, S., Chaolun, L. & Xin, S. Distinctive mitochondrial genome of Calanoid copepod Calanus sinicus with multiple large non-coding regions and reshuffled gene order: Useful molecular markers for phylogenetic and population studies. *BMC Genomics* **12**, 73, doi:10.1186/1471-2164-12-73 (2011).
126 Jondelius, U., Wallberg, A., Hooge, M. & Raikova, O. I. How the Worm Got its Pharynx: Phylogeny, Classification and Bayesian Assessment of Character Evolution in Acoela. *Systematic Biology* **60**, 845-871, doi:10.1093/sysbio/syr073 (2011).
127 Bourlat, S. J., Rota-Stabelli, O., Lanfear, R. & Telford, M. J. The mitochondrial genome structure of *Xenoturbella bocki*(phylum Xenoturbellida) is ancestral within the deuterostomes. *BMC Evolutionary Biology* **9**, 107, doi:10.1186/1471-2148-9-107 (2009).
128 Rodewald, R. & Karnovsky, M. J. Porous substructure of the glomerular slit diaphragm in the rat and mouse. *The Journal of Cell Biology* **60**, 423-433 (1974).
129 Haraldsson, B., Nyström, J. & Deen, W. M. Properties of the Glomerular Barrier and Mechanisms of Proteinuria. *Physiological Reviews* **88**, 451-487, doi:10.1152/physrev.00055.2006 (2008).

130 Weavers, H. *et al.* The insect nephrocyte is a podocyte-like cell with a filtration slit diaphragm. *Nature* **457**, 322-326, doi:http://www.nature.com/nature/journal/v457/n7227/suppinfo/nature07526_S1.html (2009).
131 Denholm, B. & Skaer, H. Bringing together components of the fly renal system. *Current Opinion in Genetics & Development* **19**, 526-532, doi:http://dx.doi.org/10.1016/j.gde.2009.08.006 (2009).
132 Na, J. & Cagan, R. The Drosophila Nephrocyte: Back on Stage. *Journal of the American Society of Nephrology* **24**, 161-163, doi:10.1681/asn.2012121227 (2013).
133 Crossley, A. C. The ultrastructure and function of pericardial cells and other nephrocytes in an insect: *Calliphora erythrocephala*. *Tissue and Cell* **4**, 529-560, doi:http://dx.doi.org/10.1016/S0040-8166(72)80029-6 (1972).
134 Liu, G. *et al.* Neph1 and nephrin interaction in the slit diaphragm is an important determinant of glomerular permeability. *The Journal of Clinical Investigation* **112**, 209-221, doi:10.1172/JCI18242 (2003).
135 Gerke, P., Huber, T. B., Sellin, L., Benzing, T. & Walz, G. Homodimerization and Heterodimerization of the Glomerular Podocyte Proteins Nephrin and NEPH1. *Journal of the American Society of Nephrology* **14**, 918-926, doi:10.1097/01.asn.0000057853.05686.89 (2003).
136 Barletta, G.-M., Kovari, I. A., Verma, R. K., Kerjaschki, D. & Holzman, L. B. Nephrin and Neph1 Co-localize at the Podocyte Foot Process Intercellular Junction and Form cis Hetero-oligomers. *Journal of Biological Chemistry* **278**, 19266-19271, doi:10.1074/jbc.M301279200 (2003).
137 Ruotsalainen, V. *et al.* Nephrin is specifically located at the slit diaphragm of glomerular podocytes. *Proceedings of the National Academy of Sciences* **96**, 7962-7967, doi:10.1073/pnas.96.14.7962 (1999).
138 Tepass, U. & Hartenstein, V. The Development of Cellular Junctions in the Drosophila Embryo. *Developmental Biology* **161**, 563-596, doi:http://dx.doi.org/10.1006/dbio.1994.1054 (1994).
139 Zhuang, S. *et al.* Sns and Kirre, the Drosophila orthologs of Nephrin and Neph1, direct adhesion, fusion and formation of a slit diaphragm-like structure in insect nephrocytes. *Development* **136**, 2335-2344, doi:10.1242/dev.031609 (2009).
140 Huber, T. B. *et al.* The Carboxyl Terminus of Neph Family Members Binds to the PDZ Domain Protein Zonula Occludens-1. *Journal of Biological Chemistry* **278**, 13417-13421, doi:10.1074/jbc.C200678200 (2003).
141 Shih, N.-Y. *et al.* CD2AP Localizes to the Slit Diaphragm and Binds to Nephrin via a Novel C-Terminal Domain. *The American Journal of Pathology* **159**, 2303-2308, doi:http://dx.doi.org/10.1016/S0002-9440(10)63080-5 (2001).
142 Boute, N. *et al.* NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. *Nature Genetics* **24**, 349-354,

doi:http://www.nature.com/ng/journal/v24/n4/suppinfo/ng0400_349_S1.html (2000).
143 Schwarz, K. *et al.* Podocin, a raft-associated component of the glomerular slit diaphragm, interacts with CD2AP and nephrin. *The Journal of Clinical Investigation* **108**, 1621-1629, doi:10.1172/JCI12849 (2001).
144 Roselli, S. *et al.* Podocin Localizes in the Kidney to the Slit Diaphragm Area. *The American Journal of Pathology* **160**, 131-139, doi:http://dx.doi.org/10.1016/S0002-9440(10)64357-X (2002).
145 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652, doi:http://www.nature.com/nbt/journal/v29/n7/abs/nbt.1883.html - supplementary-information (2011).
146 Putaala, H., Soininen, R., Kilpeläinen, P., Wartiovaara, J. & Tryggvason, K. The murine nephrin gene is specifically expressed in kidney, brain and pancreas: inactivation of the gene leads to massive proteinuria and neonatal death. *Human Molecular Genetics* **10**, 1-8, doi:10.1093/hmg/10.1.1 (2001).
147 Neumann-Haefelin, E. *et al.* A model organism approach: defining the role of Neph proteins as regulators of neuron and kidney morphogenesis. *Human Molecular Genetics* **19**, 2347-2359, doi:10.1093/hmg/ddq108 (2010).
148 Sellin, L. *et al.* NEPH1 defines a novel family of podocin-interacting proteins. *The FASEB Journal*, doi:10.1096/fj.02-0242fje (2002).
149 Stoeckli, E. T. in *Cell Adhesion* (eds Jürgen Behrens & W. James Nelson) 373-401 (Springer Berlin Heidelberg, 2004).
150 Aplin, A. E., Howe, A., Alahari, S. K. & Juliano, R. L. Signal transduction and signal modulation by cell adhesion receptors: the role of integrins, cadherins, immunoglobulin-cell adhesion molecules, and selectins. *Pharmacological Reviews* **50**, 197-263 (1998).
151 Putalla, H., Sainio, K., Sariola, H. & Tryggvason, K. Primary Structure of Mouse and Rat Nephrin cDNA and Structure and Expression of the Mouse Gene. *Journal of the American Society of Nephrology* **11**, 991-1001 (2000).
152 Donoviel, D. B. *et al.* Proteinuria and Perinatal Lethality in Mice Lacking NEPH1, a Novel Protein with Homology to NEPHRIN. *Molecular and Cellular Biology* **21**, 4829-4836, doi:10.1128/mcb.21.14.4829-4836.2001 (2001).
153 Völker, L. A. *et al.* Comparative analysis of Neph gene expression in mouse and chicken development. *Histochemistry and Cell Biology* **137**, 355-366, doi:10.1007/s00418-011-0903-2 (2012).
154 Wanner, N. *et al.* Functional and Spatial Analysis of <italic>C. elegans</italic> SYG-1 and SYG-2, Orthologs of the Neph/Nephrin Cell Adhesion Module Directing Selective Synaptogenesis. *PLoS ONE* **6**, e23598, doi:10.1371/journal.pone.0023598 (2011).
155 Patrakka, J. & Tryggvason, K. Nephrin – a unique structural and signaling protein of the kidney filter. *Trends in Molecular Medicine* **13**, 396-403, doi:http://dx.doi.org/10.1016/j.molmed.2007.06.006 (2007).
156 Beltcheva, O. *et al.* Alternatively Used Promoters and Distinct Elements Direct Tissue-Specific Expression of Nephrin. *Journal of the*

American Society of Nephrology **14**, 352-358, doi:10.1097/01.asn.0000043081.65110.c4 (2003).
157 Lapatsina, L., Brand, J., Poole, K., Daumke, O. & Lewin, G. R. Stomatin-domain proteins. *European Journal of Cell Biology* **91**, 240-245, doi:http://dx.doi.org/10.1016/j.ejcb.2011.01.018 (2012).
158 Andrikou, C., Thiel, D., Ruiz-Santiesteban, J. A. & Hejnol, A. Excretion Through Digestive Tissues Predates The Evolution Of Excretory Organs. *bioRxiv* (2017).
159 Green, J. B. & Young, J. P. W. Slipins: ancient origin, duplication and diversification of the stomatin protein family. *BMC Evolutionary Biology* **8**, 44, doi:10.1186/1471-2148-8-44 (2008).
160 Roselli, S. *et al.* Podocin localizes in the kidney to the slit diaphragm area. *Am J Pathol* **160**, doi:10.1016/s0002-9440(10)64357-x (2002).
161 Huber, T. B. *et al.* Podocin and MEC-2 bind cholesterol to regulate the activity of associated ion channels. *Proceedings of the National Academy of Sciences* **103**, 17079-17086, doi:10.1073/pnas.0607465103 (2006).
162 Tossidou, I. *et al.* CD2AP regulates SUMOylation of CIN85 in podocytes. *Molecular and Cellular Biology* **32**, 1068-1079, doi:10.1128/MCB.06106-11 (2012).
163 Johnson, R. I., Seppa, M. J. & Cagan, R. L. The *Drosophila* CD2AP/CIN85 orthologue Cindr regulates junctions and cytoskeleton dynamics during tissue patterning. *The Journal of Cell Biology* **180**, 1191-1204, doi:10.1083/jcb.200706108 (2008).
164 de Mendoza, A., Suga, H. & Ruiz-Trillo, I. Evolution of the MAGUK protein gene family in premetazoan lineages. *BMC Evolutionary Biology* **10**, 93, doi:10.1186/1471-2148-10-93 (2010).
165 Pan, L., Chen, J., Yu, J., Yu, H. & Zhang, M. The Structure of the PDZ3-SH3-GuK Tandem of ZO-1 Protein Suggests a Supramodular Organization of the Membrane-associated Guanylate Kinase (MAGUK) Family Scaffold Protein Core. *The Journal of Biological Chemistry* **286**, 40069-40074, doi:10.1074/jbc.C111.293084 (2011).
166 Itoh, M. *et al.* The Structural and Functional Organization of the Podocyte Filtration Slits Is Regulated by Tjp1/ZO-1. *PLOS ONE* **9**, e106621, doi:10.1371/journal.pone.0106621 (2014).
167 Bauer, H., Zweimueller-Mayer, J., Steinbacher, P., Lametschwandtner, A. & Bauer, H. C. The Dual Role of Zonula Occludens (ZO) Proteins. *Journal of Biomedicine and Biotechnology* **2010**, doi:10.1155/2010/402593 (2010).
168 Semmler, H., Chiodin, M., Bailly, X., Martinez, P. & Wanninger, A. Steps towards a centralized nervous system in basal bilaterians: Insights from neurogenesis of the acoel *Symsagittifera roscoffensis*. *Development, Growth & Differentiation* **52**, 701-713, doi:10.1111/j.1440-169X.2010.01207.x (2010).
169 Bery, A., Cardona, A., Martinez, P. & Hartenstein, V. Structure of the central nervous system of a juvenile acoel, *Symsagittifera roscoffensis*. *Development Genes and Evolution* **220**, 61-76, doi:10.1007/s00427-010-0328-2 (2010).
170 Chiodin, M., Achatz, J. G., Wanninger, A. & Martinez, P. Molecular Architecture of Muscles in an Acoel and Its Evolutionary Implications.

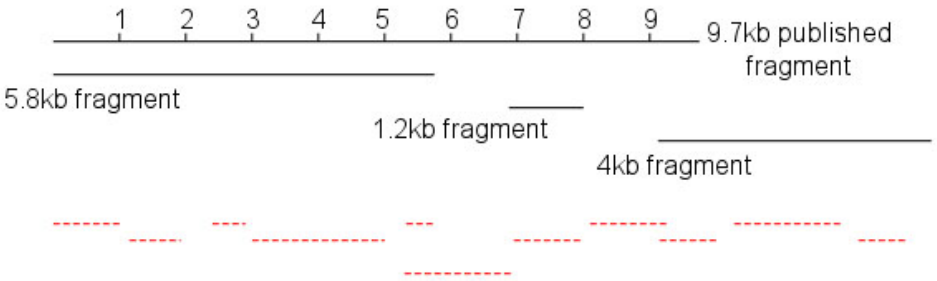
Journal of experimental zoology. Part B, Molecular and developmental evolution **316B**, doi:10.1002/jez.b.21416 (2011).
171 Dupont, S., Moya, A. & Bailly, X. Stable Photosymbiotic Relationship under CO2-Induced Acidification in the Acoel Worm *Symsagittifera roscoffensis*. *PLOS ONE* **7**, e29568, doi:10.1371/journal.pone.0029568 (2012).
172 Bailly, X. *et al.* The chimerical and multifaceted marine acoel *Symsagittifera roscoffensis*: from photosymbiosis to brain regeneration. *Frontiers in Microbiology* **5**, doi:10.3389/fmicb.2014.00498 (2014).
173 Li, M. *et al.* Nephrin expression in adult rodent central nervous system and its interaction with glutamate receptors. *The Journal of Pathology* **225**, 118-128, doi:10.1002/path.2923 (2011).
174 Yoshihara, Y., Oka, S., Ikeda, J. & Mori, K. Immunoglobulin superfamily molecules in the nervous system. *Neuroscience Research* **10**, 83-105, doi:http://dx.doi.org/10.1016/0168-0102(91)90033-U (1991).
175 Mannsfeldt, A. G., Carroll, P., Stucky, C. L. & Lewin, G. R. Stomatin, a MEC-2 Like Protein, Is Expressed by Mammalian Sensory Neurons. *Molecular and Cellular Neuroscience* **13**, 391-404, doi:http://dx.doi.org/10.1006/mcne.1999.0761 (1999).
176 Zhang, S. *et al.* MEC-2 Is Recruited to the Putative Mechanosensory Complex in *C. elegans* Touch Receptor Neurons through Its Stomatin-like Domain. *Current Biology* **14**, 1888-1896, doi:10.1016/j.cub.2004.10.030.
177 De Mulder, K. *et al.* Characterization of the stem cell system of the acoel *Isodiametra pulchra*. *BMC Developmental Biology* **9**, doi:10.1186/1471-213x-9-69 (2009).
178 Westblad, E. *Xenoturbella bocki* n.g, n.sp, a peculiar, primitive turbellarian type. *Arkiv för zoologi* **1** (1949).
179 Robinow, S. & White, K. Characterization and spatial distribution of the ELAV protein during *Drosophila melanogaster* development. *Journal of Neurobiology* **22**, 443-461, doi:10.1002/neu.480220503 (1991).
180 Berger, C., Renner, S., Lüer, K. & Technau, G. M. The commonly used marker ELAV is transiently expressed in neuroblasts and glial cells in the *Drosophila* embryonic CNS. *Developmental Dynamics* **236**, 3562-3568, doi:10.1002/dvdy.21372 (2007).
181 Marlow, H. Q., Srivastava, M., Matus, D. Q., Rokhsar, D. & Martindale, M. Q. Anatomy and development of the nervous system of *Nematostella vectensis*, an anthozoan cnidarian. *Developmental Neurobiology* **69**, 235-254, doi:10.1002/dneu.20698 (2009).
182 Etchevers, H. C., Vincent, C., Le Douarin, N. M. & Couly, G. F. The cephalic neural crest provides pericytes and smooth muscle cells to all blood vessels of the face and forebrain. *Development* **128**, 1059 (2001).
183 Lowe, C. J. *et al.* Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113**, 853-865, doi:S0092867403004690 [pii] (2003).

184 La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e519, doi:https://doi.org/10.1016/j.cell.2016.09.027 (2016).
185 Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138 (2015).
186 Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323.e1330, doi:https://doi.org/10.1016/j.cell.2016.07.054 (2016).
187 Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661 (2017).
188 Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nature reviews. Genetics* **9**, 868-882 (2008).
189 Willmer, E. N. *Cytology and Evolution*. (Academic Press, 1970).
190 Arendt, D. Genes and homology in nervous system evolution: comparing gene functions, expression patterns, and cell type molecular fingerprints. *Theory in Biosciences* **124**, 185-197, doi:S1431-7613(05)00040-6 [pii]
10.1016/j.thbio.2005.08.002 (2005).
191 Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491-1498, doi:10.1101/gr.190595.115 (2015).
192 Sebé-Pedrós, A. *et al.* Cnidarian cell type diversity revealed by whole-organism single-cell RNA-seq analysis.
193 Liang, X. *et al.* Isl1 Is required for multiple aspects of motor neuron development. *Molecular and cellular neurosciences* **47**, 215-222, doi:10.1016/j.mcn.2011.04.007 (2011).
194 Ehlers, U. & Sopott Ehlers, B. Ultrastructure of the subepidermal musculature of *Xenoturbella bocki*, the adelphotaxon of the Bilateria. *Zoomorphology* **117**, doi:10.1007/s004350050032 (1997).
195 Brunet, T. *et al.* The evolutionary origin of bilaterian smooth and striated myocytes. *eLife* **5**, e19607, doi:10.7554/eLife.19607 (2016).
196 Südhof, T. C. Synaptotagmins: Why So Many? *Journal of Biological Chemistry* **277**, 7629-7632, doi:10.1074/jbc.R100052200 (2002).
197 Lauri, A., Bertucci, P. & Arendt, D. Neurotrophin, p75, and Trk Signaling Module in the Developing Nervous System of the Marine Annelid *Platynereis dumerilii*. *BioMed Research International* **2016**, 2456062, doi:10.1155/2016/2456062 (2016).
198 Kim, N., Park, C., Jeong, Y. & Song, M.-R. Functional Diversification of Motor Neuron-specific Isl1 Enhancers during Evolution. *PLOS Genetics* **11**, e1005560, doi:10.1371/journal.pgen.1005560 (2015).
199 Marlow, H. & Arendt, D. Evolution: Ctenophore Genomes and the Origin of Neurons. *Current Biology* **24**, R757-R761, doi:http://dx.doi.org/10.1016/j.cub.2014.06.057 (2014).
200 Esposito, R. *et al.* New Insights into the Evolution of Metazoan Tyrosinase Gene Family. *PLOS ONE* **7**, e35731, doi:10.1371/journal.pone.0035731 (2012).
201 Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nature reviews. Genetics* **12**, 692-702 (2011).
202 Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages.

Proceedings of the National Academy of Sciences **106**, 7273-7280, doi:10.1073/pnas.0901808106 (2009).
203 Albà, M. M. & Castresana, J. Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. *Molecular Biology and Evolution* **22**, 598-606, doi:10.1093/molbev/msi045 (2005).
204 Pál, C., Papp, B. & Hurst, L. D. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics* **158**, 927 (2001).
205 Combs, P. A. & Eisen, M. B. Sequencing mRNA from Cryo-Sliced Drosophila Embryos to Determine Genome-Wide Spatial Patterns of Gene Expression. *PLOS ONE* **8**, e71820, doi:10.1371/journal.pone.0071820 (2013).
206 Wu, C.-C. *et al.* Spatially Resolved Genome-wide Transcriptional Profiling Identifies BMP Signaling as Essential Regulator of Zebrafish Cardiomyocyte Regeneration. *Developmental Cell* **36**, 36-49, doi:https://doi.org/10.1016/j.devcel.2015.12.010 (2016).
207 Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57-63, doi:10.1038/nrg2484 (2009).
208 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621-628, doi:http://www.nature.com/nmeth/journal/v5/n7/suppinfo/nmeth.1226_S1.html (2008).
209 Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* **30**, 777-782, doi:10.1038/nbt.2282 (2012).
210 Pan, X. *et al.* Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proceedings of the National Academy of Sciences* **110**, 594-599, doi:10.1073/pnas.1217322109 (2013).
211 Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports* **2**, doi:10.1016/j.celrep.2012.08.003 (2012).
212 Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology* **17**, 77, doi:10.1186/s13059-016-0938-8 (2016).
213 Bhargava, V., Head, S. R., Ordoukhanian, P., Mercola, M. & Subramaniam, S. Technical Variations in Low-Input RNA-seq Methodologies. **4**, 3678, doi:10.1038/srep03678 https://http://www.nature.com/articles/srep03678 - supplementary-information (2014).
214 Subkhankulova, T. & Livesey, F. J. Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single-cell level. *Genome Biology* **7**, R18-R18, doi:10.1186/gb-2006-7-3-r18 (2006).
215 Duftner, N., Larkins-Ford, J., Legendre, M. & Hofmann, H. A. Efficacy of RNA amplification is dependent on sequence characteristics: Implications for gene expression profiling using a cDNA microarray. *Genomics* **91**, 108-117, doi:https://doi.org/10.1016/j.ygeno.2007.09.004 (2008).

216 Alberti, A. *et al.* Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15**, 912, doi:10.1186/1471-2164-15-912 (2014).
217 Wadenbäck, J. *et al.* Comparison of standard exponential and linear techniques to amplify small cDNA samples for microarrays. *BMC Genomics* **6**, 61, doi:10.1186/1471-2164-6-61 (2005).
218 Holland, P. W. Evolution of homeobox genes. *Wiley Interdisciplinary Reviews: Developmental Biology* **2**, doi:10.1002/wdev.78 (2013).
219 Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* **527**, 459-465, doi:10.1038/nature16150
http://www.nature.com/nature/journal/v527/n7579/abs/nature16150.html - supplementary-information (2015).
220 Israelsson, O. Chlamydial symbionts in the enigmatic Xenoturbella (Deuterostomia). *Journal of Invertebrate Pathology* **96**, doi:10.1016/j.jip.2007.05.002 (2007).
221 Kjeldsen, K. U., Obst, M., Nakano, H., Funch, P. & Schramm, A. Two types of endosymbiotic bacetria in the enigmatic marine worm Xenoturbella bocki. *Applied and Environmental Microbiology* **76**, doi:10.1128/aem.01092-09 (2010).
222 Strong, M. J. *et al.* Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples. *PLoS Pathogens* **10**, e1004437, doi:10.1371/journal.ppat.1004437 (2014).
223 Mallo, M. & Alonso, C. R. The regulation of Hox gene expression during animal development. *Development* **140**, 3951 (2013).
224 Moreno, E., Permanyer, J. & Martinez, P. The Origin of Patterning Systems in Bilateria—Insights from the Hox and ParaHox Genes in Acoelomorpha. *Genomics, Proteomics & Bioinformatics* **9**, 65-76, doi:https://doi.org/10.1016/S1672-0229(11)60010-7 (2011).
225 Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**, doi:10.1093/nar/gki458 (2005).

Appendix 1: *P. rubra* mitochondrial contig PCR



Schematic of *P. rubra* genome assembly fragments used as a starting point for mitochondrial genome inference. PCR sequencing results bridging or covering parts of these fragments indicated by red dashed lines.

Appendix 2: Accession numbers (NCBI) for taxa used in mitochondrial phylogenetic inference

Classification	Species	Accession Number
Acoela	*Paratomella rubra*	AY228758
	Paratomella rubra (this study)	submitted
	Symsagittifera roscoffensis	NC_014578.1
	Isodiametra pulchra (this study)	submitted
	Convolutriloba longifissura	Trace Archive Library_id=CT_MM1
	Neochildia fusca	Trace Archive Library_id=NF_MM1
Annelida	*Platynereis dumerilii*	AF178678
	Urechis caupo	NC_006379
	Lumbricus terrestris	NC_001673.1
Arthropoda	*Locusta migratoria*	NC_001712
	Daphnia pulex	NC_000844
Brachiopoda	*Terebratulina retusa*	NC_000941.1
Chaetognatha	*Spadella cephaloptera*	NC_006386.1
Chordata	*Salmo salar*	LC012541.1
	Homo sapiens	NC_001807
	Lampetra fluviatilis	NC_001131
	Ornithorhynchus anatinus	NC_000891.1
	Branchiostoma floridae	NC_000834.1
	Asymmetron inferum	NC_009774.1
	Myxine glutinosa	NC_002639.1
Cnidaria	*Acropora tenuis*	NC_003522.1
	Porites porites	NC_008166
	Discosoma sp	NC_008071.1
	Nematostella sp	NC_008164
	Aurelia aurita	NC_008446.1

(continued)

Classification	Species	Accession Number
Echinodermata	*Antedon mediterranea*	NC_010692.1
	Florometra serratissima	NC_001878.1
	Strongylocentrotus pallidus	NC_009941.1
	Strongylocentrotus droebachiensis	NC_009940.1
	Strongylocentrotus purpuratus	X12631.1
	Acanthaster brevispinus	NC_007789.1
	Acanthaster planci	NC_007788.1
	Asterias amurensis	NC_006665.1
	Apostichopus japonicus	NC_012616.1
	Cucumaria miniata	NC_005929
	Ophiura albida	NC_010691.1
	Ophiopholis aculeata	NC_005334.1
Hemichordata	*Balanoglossus carnosus*	NC_001887.1
	Saccoglossus kowalevskii	NC_007438
Mollusca	*Sepia esculenta*	NC_009690.1
	Aplysia californica	NC_005827.1
	Pupa strigosa	NC_002176
	Biomphalaria tenagophila	NC_010220.1
	Siphonodentalium lobatum	NC_005840.1
Nemertodermatida	*Nemertoderma westbladi*	AY228757.1
Porifera	*Geodia neptuni*	NC_006990.1
	Axinella corrugata	NC_006894.1
	Tethya actinia	NC_006991.1
	Amphimedon queenslandica	NC_008944
Priapulida	*Priapulus caudatus*	DQ463747
Urochordata	*Ciona intestinalis*	NC_004447.2
	Doliolum nationalis	AB176541.1
Xenoturbellida	*Xenoturbella bocki*	NC_008556.1
	Xenoturbella hollandorum	NC_029218.1
	Xenoturbella monstrosa	NC_029219.1
	Xenoturbella churro	NC_029217.1
	Xenoturbella profunda	NC_029220.1

Appendix 3: Solutions

AP buffer (MABT *S. roscoffensis* protocol)

Reagent	Amount	Final Concentration
1M Tris, pH 9.5	500µl	100mM
5M NaCl	100µl	100mM
1M $MgCl_2$	250µl	50mM
10% Tween-20	50µl	0.001%
Milli-Q H2O	4.1ml	

AP buffer (PBS *S. roscoffensis* protocol)

Reagent	Amount	Final Concentration
1M NaCl	5ml	100mM
1M $MgCL_2$	2.5ml	50mM
1M Tris, pH 9.5	5ml	100mM
20% Tween-20	1.25ml	0.5%
Milli-Q H_2O	36.25ml	

Hybe buffer (*Xenoturbella* and *S. roscoffensis in situ* hybridisation on slides)

Reagent	Amount	Final Concentration
10x salt solution (pH 7.5)*	1ml	1x
Formamide	5ml	50x
50% dextran sulphate stock**	2ml	10x
yeast tRNA 10mg/ml stock	1mg/ml	1ml
50x Denhardt's solution	200µl	1x
DEPC-treated H_2O	to 10ml total (800µl)	

*a stock of 10x salt solution was prepared as below:

Reagent	Amount	Final Concentration
NaCl	23.38g	2M
Tris HCl	2.81g	89mM
Tris base	0.265g	11mM (total Tris =0.1M)
$NaH_2PO_4.H_2O$	1.38g	50mM
Na_2HPO_4	1.42g	50mM (total PO_4=0.1M)
0.5M EDTA	20ml	50mM
DEPC-treated H_2O	to 200ml total	

**a stock of dextran sulphate stock was prepared by dissolving 5g dextran sulphate in 10ml DEPC-treated H_2O on a shaking platform.

Hybe buffer (MABT *S. roscoffensis* protocol)

Reagent	Amount	Final Concentration
Deionized formamide	500µl	50%
50% polyethylene glycol	200µl	10%
5M NaCl	120µl	600mM
2M Tris, pH 7.5	10µl	20mM
20mg/ml yeast RNA	25µl	25ug
10% Tween-20	10µl	0.001%
0.5M EDTA	10µl	5mM
50x Denhardt's solution	20µl	1x
DEPC H2O	105µl	

Hybe buffer (PBS *S. roscoffensis* protocol)

Reagent	Amount	Final Concentration
Formamide	20ml	50%
20x SSC (pH 4.5)*	10ml	5x
20 mg/ml heparin	100µl	50ug/ml
20% Tween-20	200µl	0.1%
20% SDS	2ml	1%
10mg/ml SS DNA	200µl	100ug/ml
DEPC-treated dH_2O	7.8ml	

*20x SSC solution prepared as:

Reagent	Amount
NaCl	175.3g
Sodium citrate dehydrate	88.2g
H_2O	to 1L
Adjusted to pH 5 with citric acid	

Maleic acid buffer (MABT *S. roscoffensis* protocol)

Reagent	Amount	Final Concentration
Maleic Acid solution (pH 7.5)	5ml	100mM
NaCl solution (1M)	1.5ml	150mM
Tween-20	50µl	0.1%
DEPC H2O	43.45ml	

Maleic acid buffer (PBS *S. roscoffensis* protocol)

Reagent	Amount	Final Concentration
Maleic Acid	2.902g	100mM
NaCl	2.1915g	150mM

Milli-Q H_2O added to a total volume of 250ml. Final pH was adjusted to 7.5 using NaOH, and the solution autoclaved before use.

Maleic acid buffer (*Xenoturbella* and *S. roscoffensis in situ* hybridisation on slides)

To prepare the 10x MAB stock, 116g of maleic acid was dissolved in 700ml Milli-Q H_2O and adjusted to pH 6.5 using NaOH pellets, added in small batches. Tris base was used to adjust the pH of the solution to 7.5, to which 87g NaCl was added, and made up to 1L with Milli-Q H_2O.

NMTM

Reagent	Amount	Final Concentration
5M NaCl	2ml	100mM
1M Tris pH 9.5	10ml	100mM
1M $MgCl_2$	5ml	5mM
Tween-20	100µl	0.1%
Milli-Q H_2O	to 100ml	

RNA Fragmentation Buffer

Reagent	Amount	Final Concentration
Tris Acetate, pH 8.1*	4ml	200mM
MgOAc	0.64g	150mM
KOAc	0.98g	500mM
DEPC-H_2O	to 20ml	

Fragmentation buffer mixed thoroughly on magnetic stirrer and filtered using a 0.22µm syringe filter before use.

*Tris Acetate prepared as 24.2g Trizma base dissolved in 150ml DEPC- H_2O. pH adjusted to 8.1 using glacial acetic acid, and DEPC- H_2O added to a final volume of 200ml.

SDS lysis buffer

Reagent	Amount	Final Concentration
Tris-HCl	100ml	100mM
0.5M EDTA pH8	100ml	50mM
10% SDS*	125ml	1.25%
DEPC-H_2O	675ml	

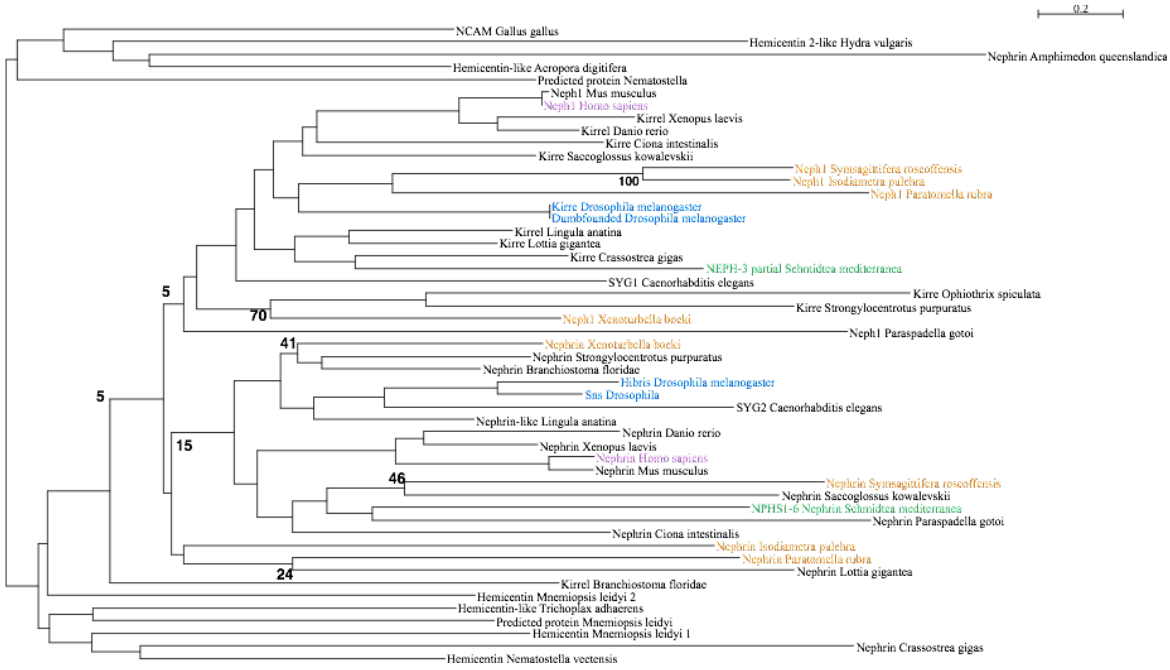
*10% SDS prepared as stock solution of 100g SDS dissolved in DEPC-H_2O to a final volume of 1L.

Appendix 4: RNA probe primer sequences and lengths

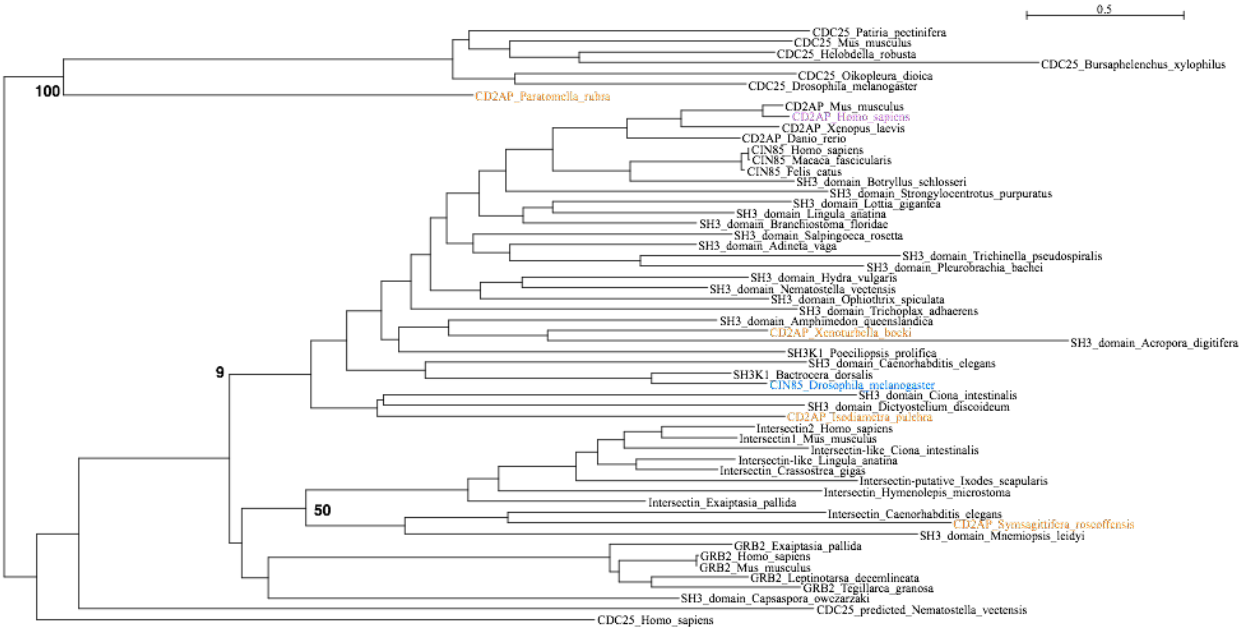
Gene	Forward Primer	Reverse Primer
SrTroponin I	TTCCGTTCAGCTTCTGGTCT	TTCGTTTGGCTGAGTGTTTG
SrNeph1	TGTGCAAGGCCAAGAATGAC	TCAACTCACGCTTCTCCACT
SrNephrin	AAACAGACTGCAAACCCACC	CAACATACCCGCGTGATCTG
SrPodocin-like	CCAGTGGCAGTGTTCATGAC	GCTCAACGTCTGCAGGTATC
XbElav	AGGCTGTGTCTACCCTGAAC	ATAGAACAGGACGGTGTGGG
XbNeph1	CACGGTGAGTTTGGTGTGTC	ACCACACCCATTCCATGTTT
XbNephrin	GTCAGAGTTGGAGTGGTCGA	AGCAGTCTCACCCTCCTCTA
XbPodocin-like	TGCTTTATTTGGCAAGAGCA	AGGGGTCAGTGTTTGTGAGG
XbTroponin C	GTACATACGCATCTCCGCTCATC	GCGTAATACGACTCACTATAGGAATACAT TCCAACGGTCCTCCTC
XbTroponin T	ATAGCTGAGTCGCGCATACACAT	GCGTAATACGACTCACTATAGGAATTCAG CCGCTCGTTCCTTAG
Innexin	CCACCTGTGGTATCAGTGGATTC	GCGTAATACGACTCACTATAGGCGATAGG AATCTCCGAACTGTCA
XbChymotrypsin	ATCGCAATCGCAGAATTCAA	GCGTAATACGACTCACTATAGGTTAGGCC TAGTCCAGTGACACAT
XbTyrosinase-like	GAAGGAGTATCACAGCGATGGAA	GCGTAATACGACTCACTATAGGGGTGCAA GGAACTGATTATGGTG
XbNeural-specific	ACTTCAACACACGACGCAATCTT	GCGTAATACGACTCACTATAGGGCATATG CACGTGAACACGAA

Gene	Length of probe (bp)	Approximate location of probe sequence
SrNeph1	768	ORF, whole region, beginning ~400bp from 5' start codon
SrNephrin	1051	ORF, whole region, beginning ~100bp from 5' start codon
SrPodocin-like	633	ORF, 3' end, beginning ~2000bp from 5' start codon
XbNeph1	926	ORF, whole region, beginning ~250bp from 5' start codon
XbNephrin	1351	ORF, whole region, beginning ~180 bp from 5' start codon
XbPodocin-like	1062	Mostly ORF: probe sequence starts in the 5' UTR region, ~150bp before the start codon.

Appendix 5: Bootstrap support for CAM (Neph1 and Nephrin) and CD2AP proteins.

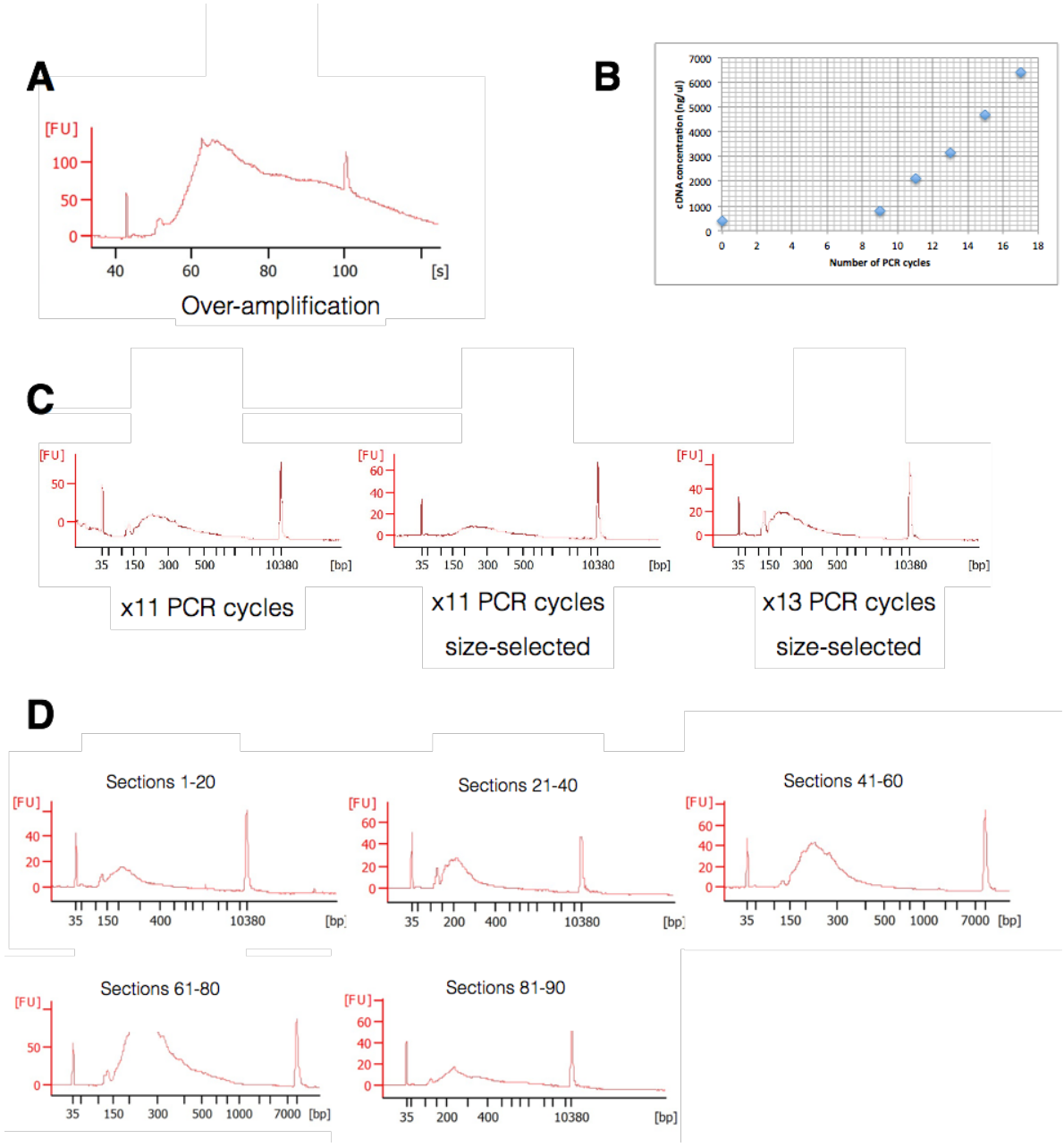


Maximum likelihood analysis of the CAM family proteins Neph1 and Nephrin. Xenacoelomorpha sequences shown in orange; *D. melanogaster* sequences in blue; *H. sapiens* sequences in purple; *S. mediterranea* in green. Phylogenetic inference carried out using RAxML. Bootstrap support values at relevant nodes. Branch length value shows number of substitutions per site.



Maximum likelihood analysis of the SH3-domain CD2AP protein and related outgroups. Xenacoelomorpha sequences shown in orange; *D. melanogaster* sequences in blue; *H. sapiens* sequences in purple. Phylogenetic inference carried out using RAxML. Bootstrap support values at relevant nodes. Branch length value shows number of substitutions per site.

Appendix 6: Protocol refinement for Tomoseq

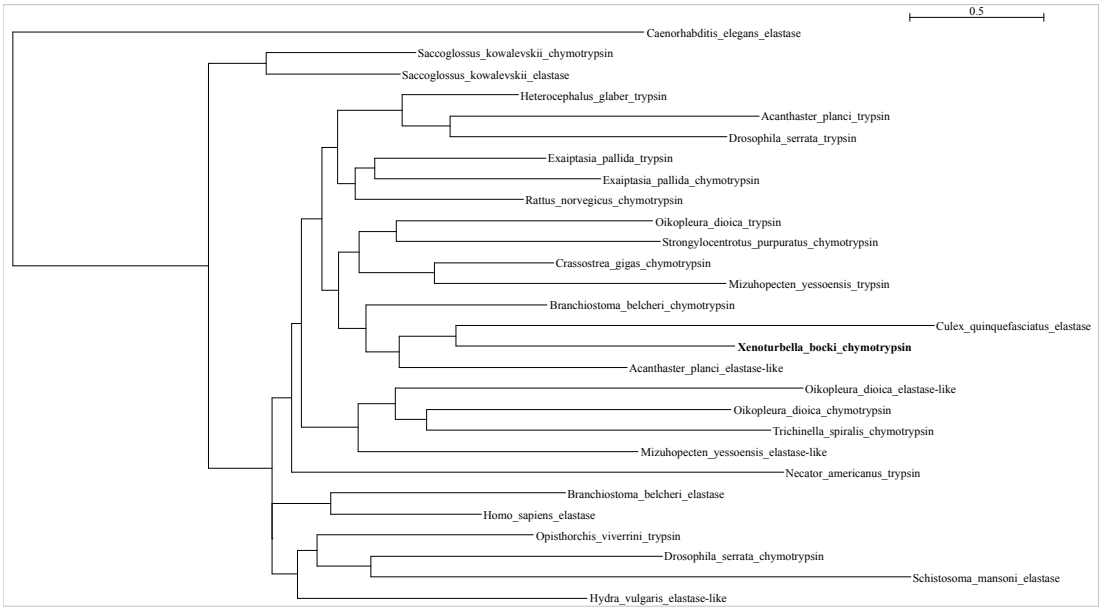


(A) Bioanalyzer curve resulting from overamplification of RNA; (B) Test PCR cycles to establish linear amplification; (C) Comparison of 11 cycle vs. 13 cycle PCR library amplification; (D) Bioanalyzer curves for pooled *Xenoturbella* 90-section Tomoseq.

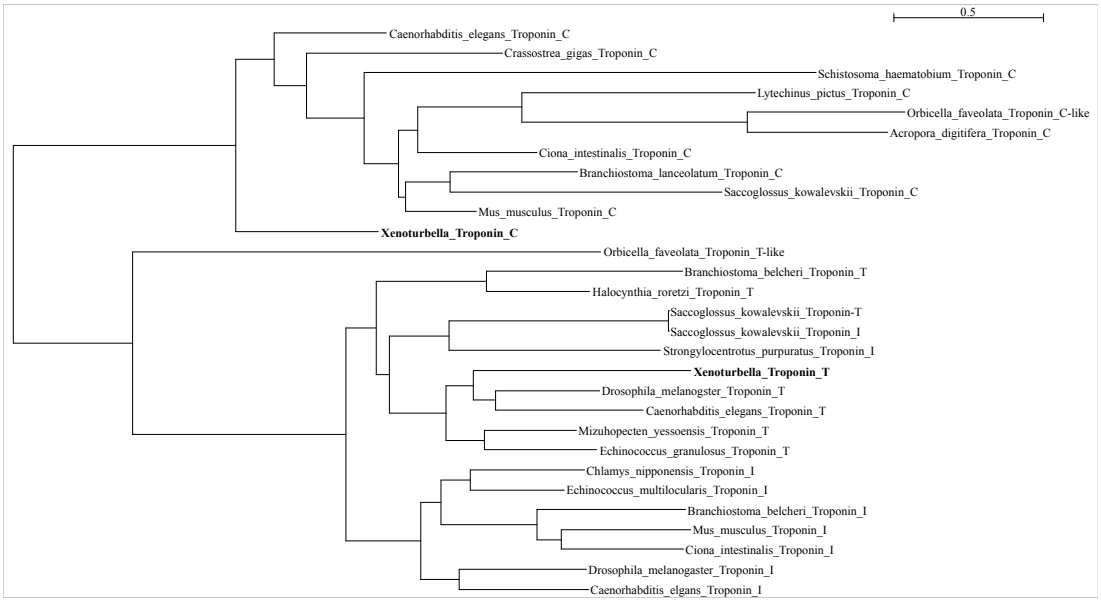
Appendix 7: CelSeq2 primer sequences

Primer	Sequence
1	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNAGACTCTTTTTTTTTTTTTTTTTTTTTTTV
2	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNAGCTAGTTTTTTTTTTTTTTTTTTTTTTTV
4	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNAGCTTCTTTTTTTTTTTTTTTTTTTTTTTV
5	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNCATGAGTTTTTTTTTTTTTTTTTTTTTTTV
9	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNCAGATCTTTTTTTTTTTTTTTTTTTTTTTV
10	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNTCACAGTTTTTTTTTTTTTTTTTTTTTTTV
11	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNAGGATCTTTTTTTTTTTTTTTTTTTTTTTV
14	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNTCCTAGTTTTTTTTTTTTTTTTTTTTTTTV
17	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNTCGAAGTTTTTTTTTTTTTTTTTTTTTTTV
20	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNGTACAGTTTTTTTTTTTTTTTTTTTTTTTV
23	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNGTCTAGTTTTTTTTTTTTTTTTTTTTTTTV
25	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNGTTGCATTTTTTTTTTTTTTTTTTTTTTTV
26	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNGTGACATTTTTTTTTTTTTTTTTTTTTTTV
28	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNACAGTGTTTTTTTTTTTTTTTTTTTTTTTV
29	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNACCATGTTTTTTTTTTTTTTTTTTTTTTTV
31	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNACTCGATTTTTTTTTTTTTTTTTTTTTTTV
32	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNACGTACTTTTTTTTTTTTTTTTTTTTTTTV
35	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNCTAGACTTTTTTTTTTTTTTTTTTTTTTTV
40	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNCTTCGATTTTTTTTTTTTTTTTTTTTTTTV
46	GCCGGTAATACGACTCACTATAGGGAGTTCTACAGTCCGACGATCNNNNNNTGCAGATTTTTTTTTTTTTTTTTTTTTTTV
RPI2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
RPI5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
RPI6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
RPI9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
RPI10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
RPI11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

Appendix 8: Orthology assignment of meta-cluster specific genes identified in the single cell sequencing protocol.



Chymotrypsin-like



Troponin

SCIENTIFIC REPORTS

OPEN

The mitochondrial genomes of the acoelomorph worms *Paratomella rubra, Isodiametra pulchra and Archaphanostoma ylvae*

Received: 27 January 2017
Accepted: 31 March 2017
Published online: 12 May 2017

Helen E. Robertson[1], François Lapraz[1,2], Bernhard Egger[1,3], Maximilian J. Telford[1] & Philipp H. Schiffer[1]

Acoels are small, ubiquitous - but understudied - marine worms with a very simple body plan. Their internal phylogeny is still not fully resolved, and the position of their proposed phylum Xenacoelomorpha remains debated. Here we describe mitochondrial genome sequences from the acoels *Paratomella rubra and Isodiametra pulchra*, and the complete mitochondrial genome of the acoel *Archaphanostoma ylvae*. *The P. rubra and A. ylvae* sequences are typical for metazoans in size and gene content. The larger *I. pulchra* mitochondrial genome contains both ribosomal genes, 21 tRNAs, but only 11 protein-coding genes. We find evidence suggesting a duplicated sequence in the *I. pulchra* mitochondrial genome. The *P. rubra, I. pulchra and A. ylvae* mitochondria have a unique genome organisation in comparison to other metazoan mitochondrial genomes. We found a large degree of protein-coding gene and tRNA overlap with little non-coding sequence in the compact *P. rubra* genome. Conversely, the *A. ylvae* and *I. pulchra* genomes have many long non-coding sequences between genes, likely driving genome size expansion in the latter. Phylogenetic trees inferred from mitochondrial genes retrieve Xenacoelomorpha as an early branching taxon in the deuterostomes. Sequence divergence analysis between *P. rubra* sampled in England and Spain indicates cryptic diversity.

Acoel flatworms are small, soft-bodied, unsegmented, marine animals lacking a gut epithelium, coelomic cavity, and anus. Instead, they typically possess a ventral mouth opening, and a simple syncytial digestive system[1]. Due primarily to the common attributes of acoelomate body and the absence of a through gut, Acoela were traditionally grouped as an order within the Platyhelminthes. The first molecular systematic studies on these animals using small subunit (SSU) ribosomal RNA gene sequences revealed that the Acoelomorpha are in fact a distinct lineage, quite separate from the main clade of the Platyhelminthes (Rhabditophora and Catenulida)[2–4]. Instead, these initial molecular studies supported a position of the Acoelomorpha diverging prior to the protostome/deuterostome common ancestor. More recently, the Acoelomorpha have been linked to the similarly simple marine worm *Xenoturbella* in the new phylum Xenacoelomorpha, making sense of their shared simple body plan and other shared morphological characters, such as unusual ciliary ultrastructure[5] and their simple basiepidermal nervous system[6]. Despite considerable efforts, the position of Xenacoelomorpha within the Metazoa remains unresolved, with alternative lines of evidence placing them either as the sister group to the remaining Bilateria (protostomes and deuterostomes)[7,8], or as a phylum within the deuterostomes[9]. A better understanding of acoel phylogeny and evolution is therefore integral to answering central questions concerning the evolution of Bilateria and its subtaxa. To this end more genomic data are needed.

Metazoan mitochondrial DNA (mtDNA) is a closed-circular molecule typically comprising 37 genes which are, for the most part, invariant across the Metazoa[10]. These include the two rRNAs of the mitochondrial ribosome, 22 tRNAs necessary for translation, and 13 protein-coding genes for the enzymes of oxidative phosphorylation. *atp8* is the only gene known to have been commonly lost from this complement, and this has been observed

[1]Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London, WC1E 6BT, UK. [2]Present address: CNRS/UMR 7277, institut de Biologie Valrose, iBV, Université de Nice Sophia Antipolis, Parc Valrose, Nice cedex 2, France. [3]Present address: Institute of Zoology, University of Innsbruck, Technikerstr. 25, 6020, Innsbruck, Austria. Correspondence and requests for materials should be addressed to H.E.R. (email: helen.robertson.09@ucl.ac.uk) or P.H.S. (email: philipp.schiffer@gmail.com)

in a number of independent metazoan lineages, including the acoel *Symsagittifera roscoffensis*[11]. In addition to primary sequence data, mtDNA has a number of other features which can be used for phylogenetic inference, including variations in mitochondrial genetic code[12]; a higher rate of sequence evolution than nuclear DNA[13]; changes in gene order; and changes in the secondary structure of rRNAs and tRNAs[14].

Mitochondrial gene sequences have been used extensively for phylogenetic inference. In a recent paper, Rouse *et al*. used mitochondrial protein-coding sequence data from four newly discovered species of *Xenoturbella* (*X. hollandorum*, *X. churro*, *X. monstrosa*, and *X. profunda*) to infer the internal phylogeny of the Xenoturbellida[15]. Wider phylogenetic inference including mitochondrial proteins from these species placed Xenacoelomorpha with the deuterostomes[15], corroborating previous mitochondrial phylogenetic analysis of this phylum[9,16,17].

Mitochondrial gene content is largely invariable across the Metazoa, with the order in which genes are arranged being fairly stable and conserved for up to hundreds of millions of years in some metazoan lineages. Rearrangement events, thought to occur via a model of 'duplication and deletion'[14,18], whereby a portion of the mitochondrial genome is duplicated, and the original copy of the duplicated gene subsequently deleted, are rare. The infrequency of such rearrangements, and the huge number of possible rearrangement scenarios, means that convergence on the same gene order in unrelated lineages is unlikely. Gene order is thus likely to retain evolutionary signals, with a common gene order being indicative of common ancestry and informative for the study of metazoan divergence[19]. Rearrangement of genes within the mitochondrial genome of different species can be a particularly powerful tool in the analysis of phylogenetic relationships[14] and may also indicate accelerated evolution in a taxon.

In this study, we describe the mitochondrial genomes from three species of acoel: *Paratomella rubra*, *Isodiametra pulchra* and *Archaphanostoma ylvae*. Adult specimens of all animals are approximately 1 mm in length, and, as is typical for small acoel species, they occupy the littoral and sub-littoral zones of marine ecosystems: *P. rubra* has been described across Europe and North America[20,21], and *I. pulchra* lives abundantly in the mud flats of Maine[22]. *A. ylvae* has been described from the West coast of Sweden[23]. All species move freely within the sediment by gliding on a multiciliated epidermis. First described by Rieger and Ott[21], *P. rubra* is an elongate and flattened worm belonging to the family Paratomellidae[24,25]. A 9.7 kb fragment of mitochondrial genome has previously been described from specimens of *P. rubra* collected on the Mediterranean coast of Spain[26]. *I pulchra* belongs to the family Isodiametridae; it can be maintained long-term in culture and has been used experimentally for *in situ* hybridisation, RNAi, and other molecular protocols[22,27,28]. It's use as a 'model acoel' therefore makes this species particularly valuable for investigation. *A. ylvae* also belongs to the Isodiametridae family of acoels. Originally described by Kånneby *et al*. in 2014, its *cox1* gene has been sequenced and used for classification, but no further genes from its mitochondrial genome have been sequenced[23].

Results

Genomic composition. We assembled 14,954 base pairs of the *P. rubra* mitochondrial genome, starting from three genome assembly fragments and using Sanger sequencing of additional PCR fragments (Fig. 1a). We were unable to close the circular mitochondrial genome of *P. rubra*, but our 14.9 kb sequence contains all 13 protein-coding genes, both ribosomal genes and 22 putative tRNAs. Compared to the fragment of the genome previously published we have found four additional protein-coding genes and 12 additional tRNAs[26]. All genes are found exclusively on one strand of the sequence. Allowing for overlap, protein-coding genes account for 74.79% of the genomic sequence; ribosomal genes 13.95%; tRNA genes 9.10%, and non-coding DNA just 2.04%. A 156 nucleotide-long stretch of non-coding sequence is found between *cytochrome c oxidase subunit 2* (*cox2*) and *NADH dehydrogenase subunit 1* (*nad1*).

In *P. rubra*, *trn*S2 is predicted entirely within the sequence coding for *nad1*, and also has clear deviation from the traditional 'cloverleaf' secondary structure of tRNA. In addition, three of the predicted tRNAs have minor overlaps with protein-coding genes: *trn*A with *nad3* (20 nucleotides); *trnK* with *nad4l* (18 nucleotides) and *trnS1* with *nad4* (six nucleotides); and all but five nucleotides of *trnL1* are predicted within the same sequence as *rrnL* (Fig. 1a, Table 1). With the exception of *trnT*, all predicted tRNAs have an amino-acyl acceptor stem composed of seven base pairs, and all predicted tRNAs apart from *trnT* and *trnS2* have a five base pair anticodon stem (Fig. 2). 11 tRNAs have one or two G-T mismatches in their acceptor or anticodon stems (*A*,*C*,*G*,*I*,*K*,*L1*,*L2*,*P*,*Q*,*R*,*T*). All tRNAs have a DHU arm of three or four nucleotides. The structure of the TψC arm shows greater variability, with a number of tRNAs having either a truncated stem, or the arm entirely lacking (Fig. 2).

For *I. pulchra*, we initially recovered a 13 kb contig, a 3.5 kb contig and a 19 kb contig of mitochondrial sequence from our transcriptomic data. The entire 13 kb contig and 2.4 kb of the 3.5 kb contigs were found to be perfectly matching subsets of the longer 19 kb sequence (Fig. 3). We designed several sets of PCR primers to verify the sequence between the 3′ end of the 13 kb and 5′ end of the 3.5 kb fragments found on the long 19 kb fragment (Fig. 3), however, no PCR amplification completely bridged the sequence between the 13 kb to 3.5 kb fragment. We found that the last (3′) 300 bp of the 13 kb fragment was duplicated in the opposite orientation within the end (3′) region of the 3.5 kb fragment. Although the long 19 kb fragment contained the repeated region between the 13 kb and 3.5 kb fragments, we were not able to connect sequences on both sides of the repeated region by PCR. The placement of Sanger sequencing fragments containing the repeat remained ambiguous. We are thus not confident in the assembly of the 19 kb transcriptomic sequence in this section, and therefore treat it as uncertain. Instead we focused on verifying the sequence of the two smaller fragments and on amplifying and sequencing the region lying between them. We reconfirmed the majority of the 13 kb fragment using PCR amplification and Sanger sequencing. We were also able to amplify and sequence fragments joining the 3′ end of the 3.5 kb fragment with the 5′ end of a 1.3 kb fragment containing the *rrnL* gene, which we had identified in transcriptome sequence data using blast. This contig included the duplicated region at the 3′ end of the 3.5 kb fragment (Fig. 3).

In summary, we find the *I. pulchra* mitochondrial genome to have a span of at least 18,725 base pairs (Fig. 1b) based on our PCR validation of the transcriptomic data. This covers the region from the start of the 5′ end of the

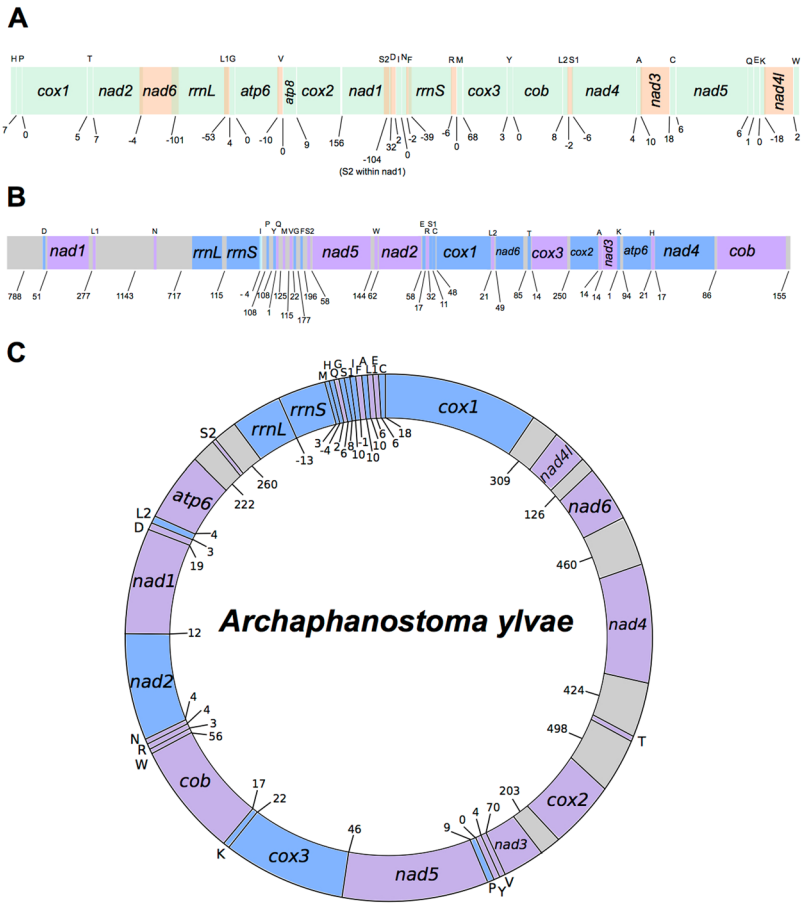


Figure 1. Overview of the mitochondrial genome sequences we resolve for *Paratomella rubra*, *Isodiametra pulchra* and *Archaphanostoma ylvae* (Xenacoelomorpha: Acoela). Genes not drawn to scale. Numbers beneath the sequences show intergenic spaces (positive values) or intergenic overlap (negative values). Protein-coding genes are denoted by three letter abbreviations; ribosomal genes by four letter abbreviations. tRNAs are shown by single uppercase letters. (**A**) *P. rubra* 14,957 base-pair long sequence. All genes found on the positive (forward) strand. Where genes, rRNAs or tRNAs are coloured orange, this is solely to demonstrate overlap with adjacent genes, rRNAs or tRNAs. (**B**) *I. pulchra* 18,725 base-pair long sequence. Genes found on the positive (forward) strand are coloured blue; genes on the negative (reverse) strand are coloured purple. Non-coding sequence shown in grey. (**C**) *A. ylvae* 16,619 nucleotide-long mitochondrial genome. Genes found on the positive (forward) strand are coloured blue; genes on the negative (reverse) strand are coloured purple. Non-coding regions greater than 100 nucleotides in length are shown in grey.

3.5 kb sequence which is linked through PCR amplicons to the 5′ end of the 13 kb sequence (including the 1.3 kb *rrnL* contig), and up to the start of the duplicated sequence at the 3′ end of the 13 kb sequence (Fig. 3). As we were not able to bridge the region between the 3′ end of the 13 kb fragment and the 5′ end of the 3.5 kb fragment with PCR, we could not confirm the validity of the duplicated sequence at this position nor fully close the circular mitochondrial genome. It is therefore likely that the entire mitochondrial genome is larger than 19 kb, and may include the duplicated sequence. The sequence we are confident on presenting contains both ribosomal genes, all tRNAs and 11 protein-coding genes. These protein-coding genes and RNAs are encoded on both the plus and minus strands. No sequences resembling either *atp8* or *nad4l* could be found in our sequence.

In the 18.7 kb sequence, protein-coding genes account for 56.66%; ribosomal genes contribute 8.15% and tRNA genes 7.77%. Compared to *the P. rubra* and *S. roscoffensis* mitochondrial genomes, intergenic space in the *I. pulchra* sequences is unusually high: non-coding DNA accounts for 22.72% of the sequences, including 14 intergenic regions of greater than 100 base pairs.

We identified all 22 expected tRNAs in the *I. pulchra* mitochondrial genome. Predicted sequences for *rrnS* and *trnI* overlap by four base pairs, but no other overlaps were found between any tRNAs or with any protein-coding genes (Fig. 1b, Table 2). All predicted tRNAs have an amino-acyl acceptor stem composed of seven base pairs and a five base pair anticodon stem, with the exception of *trnE, trnF* and *trnS2*, which have an anticodon stem composed of only four base pairs (Fig. 4). The structure of the DHU arms and TψC show greater variability, and are composed of either 3 or 4, or between 3 and 6, base pairs respectively, across the 22 tRNAs. Whilst the TψC arm is missing entirely in *trnQ*, and very truncated in *trnE*, *trnF*, *trnG* and *trnP*, more of the predicted tRNAs fit the

Feature	Strand	Start	Stop	Length (nucleotides)	Length (AA)	Start Codon	Stop Codon	Intergenic region
trnH (gtg)	+	368	426	59				7
trnP (tgg)	+	434	495	62				0
cox1	+	496	2058	1563	521	ATA	TAA	5
trnT (tgt)	+	2064	2123	60				7
nad2	+	2131	3105	975	325	ATT	TAG	−4
nad6	+	3102	3563	462	154	ATA	TAA	−101
rrnL	+	3463	4819	1357				−53
trnL1 (tag)	+	4767	4824	58				4
trnG (tcc)	+	4829	4887	59				0
atp6	+	4888	5496	609	203	ATA	TAG	−10
trnV (tac)	+	5487	5553	67				0
atp8	+	5554	5730	177	59	ATT	TA-	9
cox2	+	5740	6402	663	221	ATT	TAA	156
nad1	+	6559	7602	1044	348	ATT	TAA	−104
trnS2 (tga)	+	7499	7568	70				32
trnD (gtc)	+	7601	7662	62				2
trnI (gat)	+	7665	7728	64				0
trnN (gtt)	+	7729	7798	70				−2
trnF (gaa)	+	7797	7856	60				−39
rrnS	+	7818	8547	730				−6
trnR (tcg)	+	8542	8608	67				0
trnM (cat)	+	8609	8669	61				68
cox3	+	8738	9523	786	262	ATT	TAA	3
trnY (gta)	+	9527	9585	59				0
cob	+	9586	10668	1083	361	ATA	TAA	8
trnL2 (taa)	+	10677	10737	61				−2
trnS1 (gct)	+	10736	10799	64				−6
nad4	+	10794	12119	1326	442	ATC	TAA	4
trnA (tgc)	+	12124	12181	58				10
nad3	+	12192	12551	360	120	ATT	TAG	18
trnC (gca)	+	12570	12629	60				6
nad5	+	12636	14387	1752	584	ATA	TAG	6
trnQ (ttg)	+	14394	14449	56				1
trnE (ttc)	+	14451	14510	60				0
trnK (ttt)		14511	14576	66				−18
nad4l	+	14559	14867	309	103	ATA	TAA	2
trnW (tca)	+	14870	14933	64				

Table 1. Organisation of the Paratomella rubra 14.9 kb mitochondrial genome sequence. All genes found on the 'positive' strand.

stereotypical 'cloverleaf' secondary structure than has been found for other acoel species, including *S. roscoffensis* and *P. rubra* (Fig. 4).

The complete, closed circular mitochondrial genome of *A. ylvae* was recovered from genome sequencing data of *P. rubra* specimens collected from Yorkshire, UK. Contamination of the *P. rubra* samples was confirmed using NCBI Blast, which yielded a 99% identical sequence to the published *A. ylvae cox1*. The complete *A. ylvae* mitochondrial genome is 16,619 nucleotides in length, and contains 12 protein-coding genes, both rRNAs, and 22 predicted tRNAs (Fig. 1c, Table 3). With *cox1* at the start of the genome on the 'positive' strand, all other protein-coding genes apart from *cox3* and *nad2* are found on the 'negative' strand. Both rRNAs are found on the positive strand, and tRNAs are distributed between the two. Accounting for a small amount of overlap between genes - totalling 18 nucleotides of overlap across the whole genome - protein-coding genes make up 64.72% of the genome. tRNAs contribute 8.91%, and rRNAs 9.31%. As found for *I. pulchra*, non-coding DNA makes up a large amount of the genome, totalling 17.17%.

We identified possible sequences for all 22 mitochondrial tRNAs in the *A. ylvae* genome, although four of these (*trnE*, *trnI*, *trnK* and *trnS1*) have an e-value prediction of greater than 0.0001. All predicted secondary structures of the tRNAs in the *A. ylvae* mitochondrial genome have standard-length acceptor and anticodon stems, and the majority – with the exception of *trnK*, *L1*, *L2*, *N*, *S2* and *Y* – have a four nucleotide-long D-loop (Fig. 5). As in the other acoel mitochondrial genomes, the greatest variability is found in the TψC arm, which is truncated in *trnD*, *E*, *F*, *L1*, *P*, *V*, *W* and *Y* and missing in *trnQ* (Fig. 5).

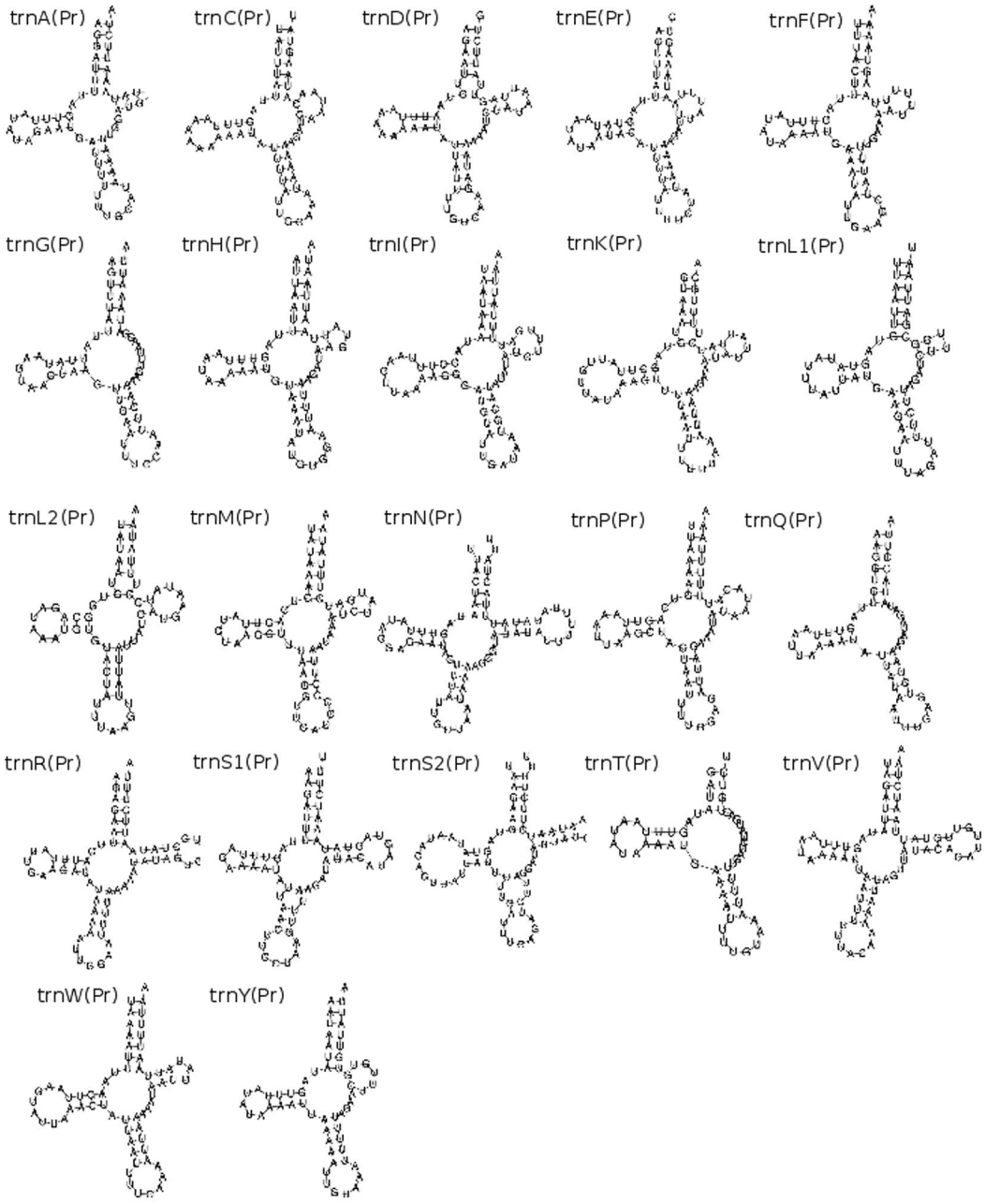


Figure 2. Predicted secondary structures of tRNAs from the mitochondrial genome sequence of *Paratomella rubra* as predicted by MiTFi in Mitos.

The *P. rubra* genome is 78.15% A + T rich, which is higher than the A + T content calculated for the *I. pulchra* genomic sequence at 67.28%, and the *A. ylvae* complete genome at 74.70%. Overall nucleotide usage on the plus strand of *P. rubra* is: A = 29.29%; T = 48.86%; C = 6.77% and G = 15.10%; GC-skew = 0.38 and absolute AT-skew = 0.25. Overall nucleotide usage for *I. pulchra* is: A = 34.04%; T = 33.24%; C = 16.45% and G = 16.27%; GC-skew = 0.006 and AT-skew = 0.012. For *A. ylvae*, A = 40.41%; T = 34.29%; C = 12.82% and G = 12.47%; GC-skew = 0.014 and AT-skew = 0.082. GC-skew and AT-skew absolute values for *P. rubra* are much higher than that of *S. roscoffensis*, although the absolute values for *I. pulchra* and *A. ylvae* are comparatively low[11]. AT-skew value for the *P. rubra* sequence is just 0.01 different from that of the published partial *P. rubra* genome, and GC-skew is slightly higher (published *P. rubra* GC-skew = 0.32)[26].

Gene order and gene arrangement. All thirteen protein-coding genes in *P. rubra* have complete initiation codons: ATA (x5) and ATT (x8). Five of the protein-coding genes previously published differ in the nucleotide sequence of their start codons: *nad2*, *atp8*, *cox2*, and *cox3* all have ATA as an initiation codon in our analysis, compared to ATT found in previous analysis[26]. Twelve of the genes have full stop codons: TAA (x9) or TAG (x3). *atp8* was found to have a truncated stop codon (TA-), which is thought to be completed during

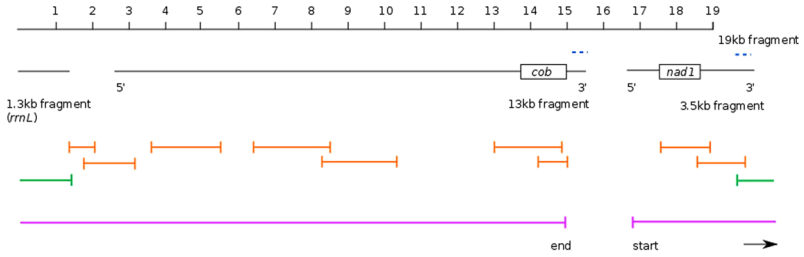


Figure 3. Overview of the initial transcriptome assembly fragments and PCR strategy for scaffolding the *Isodiametra pulchra* mitochondrial genome. 1.3 kb, 13 kb and 3.5 kb fragments aligned to a continuous 19 kb fragment, with the location of the duplicated sequence in the 13 kb and 3.5 kb fragments shown by blue dashed lines. The 'start' and 'end' regions of the 13 kb and 3.5 kb fragments are annotated by 5′ (start) and 3′ (end). The approximate location of *cob* and *nad1* protein-coding sequence are shown for reference. Reliable PCR-amplicons are shown in orange; the green PCR fragment indicates successful joining of the 3′ end of the 3.5 kb fragment to the *rrnL* fragment, including the duplicated section. The 18,725 base-pair long sequence we resolve is indicated by the pink lines, from 'start' to 'end'.

Feature	Strand	Start	Stop	Length (nucleotides)	Length (AA)	Start Codon	Stop Codon	Intergenic region
trnD (gtc)	+	789	848	60				51
nad1	−	900	1784	885	295	ATG	TAA	277
trnL1 (tag)	−	2062	2130	69				1143
trnN (gtt)	−	3274	3339	66				717
rrnL	+	4057	4657	601				115
rrnS	+	4773	5698	926				−4
trnI (gat)	+	5695	5768	74				108
trnP (tgg)	+	5877	5939	63				108
trnY (gta)	+	6048	6111	64				1
trnQ (ttg)	−	6113	6173	61				125
trnM (cat)	−	6299	6360	62				115
trnV (tac)	−	6476	6543	68				22
trnG (tcc)	+	6566	6627	62				177
trnF (gaa)	+	6805	6873	69				196
trnS2 (tga)	−	7070	7139	70				58
nad5	−	7198	8907	1710	570	ATG	TAA	144
trnW (tca)	−	9052	9118	67				62
nad2	−	9181	10233	1053	351	ATG	TAA	58
trnE (ttc)	+	10292	10355	64				17
trnR (tcg)	−	10373	10439	67				32
trnS1 (tct)	+	10472	10539	68				11
trnC (gca)	+	10551	10613	63				48
cox1	+	10662	12197	1536	512	ATA	TAG	21
trnL2 (taa)	−	12219	12286	68				49
nad6	+	12336	12811	476	159	ATG	T−	85
trnT (tgt)	+	12897	12963	67				14
cox3	−	12978	13775	798	266	ATG	TAA	250
cox2	+	14026	14640	615	205	ATA	TAA	14
trnA (tgc)	−	14655	14718	64				14
nad3	−	14733	15110	378	126	ATG	TAA	1
trnK (ttt)	+	15112	15178	67				94
atp6	+	15273	15955	683	228	ATA	TA−	21
trnH (gtg)	−	15977	16042	66				17
nad4	+	16060	17403	1344	448	ATG	TAA	86
cob	−	17490	18570	1081	361	ATA	T−	155

Table 2. Organisation of the Isodiametra pulchra 18.7 kb mitochondrial genome sequence.

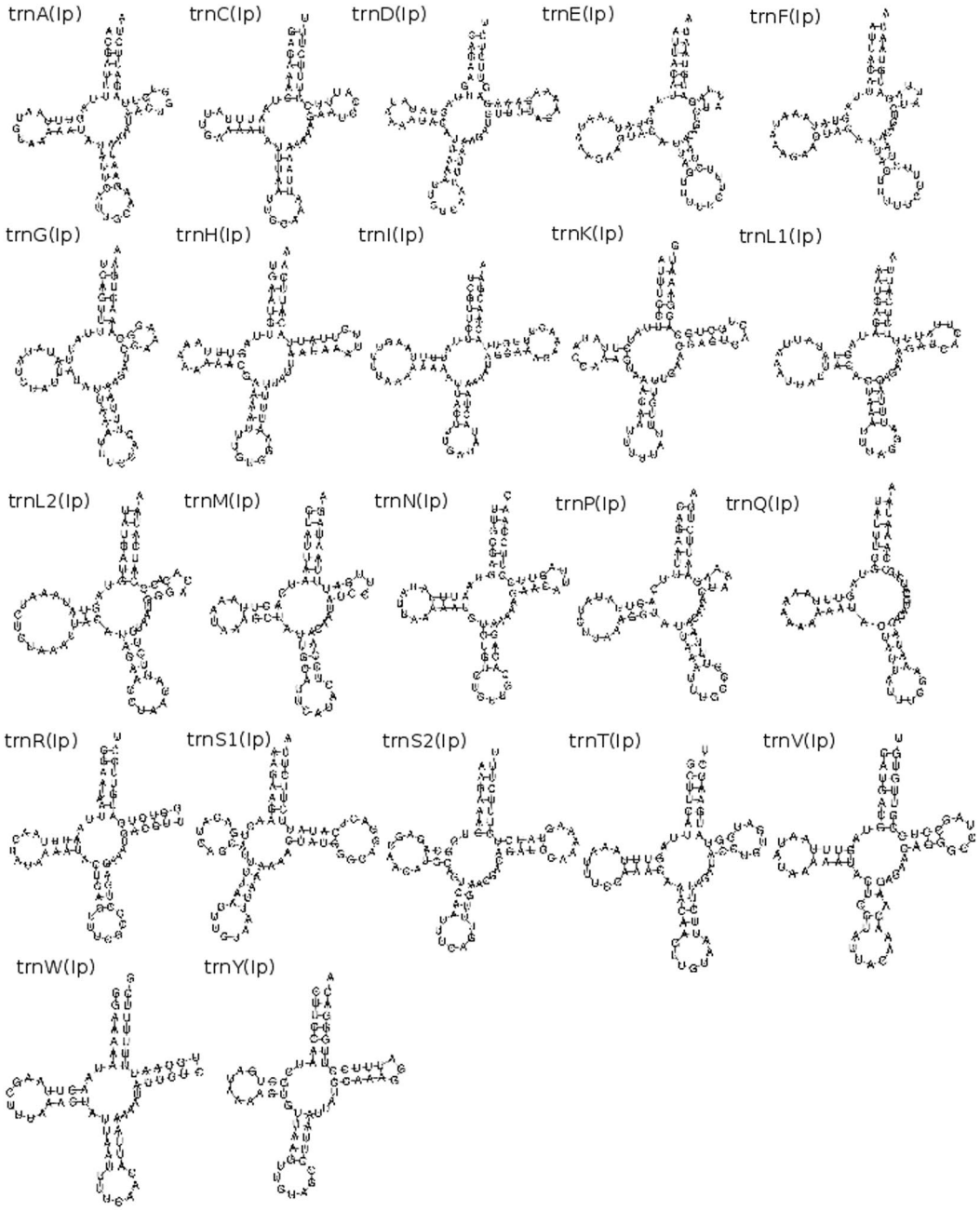


Figure 4. Predicted secondary structure of tRNAs from the mitochondrial genome sequences of and *Isodiametra pulchra* as predicted by MiTFi in Mitos.

post-transcriptional modification (Table 1). The eleven protein-coding genes found for *I. pulchra* also have full initiation codons: ATA (x4) and ATG (x7). Eight of the genes for this species have full stop codons: TAA (x7) and TAG (x1); *nad6*, *atp6* and *cob* are inferred to have truncated stop codons (Table 2). Initiation codes in *A. ylvae* are: ATA x4, ATG x7 and ATT x1; all genes have TAA as stop codons, with the exception of *nad6*, which has TAG (Table 3). As in other invertebrate mitochondrial genomes, our data indicates a deviation from the 'standard' genetic code, with ATA encoding the start codon methionine, M, instead of isoleucine, I.

We found all genes in *P. rubra* on the 'plus' strand. In *I. pulchra*, genes are distributed over the plus and minus strands, with just two 'blocks' of genes with the same transcriptional polarity clustered together (*rrnL-rrnS-trnI-trnP-trnY*; *trnS2-nad5-trnW-nad2*). Similarly, in *A. ylvae* genes are distributed across the two strands, with two clustered 'blocks' of genes and tRNAs (*nad4l-nad6-nad4-trnT-cox2-nad3-trnV-trnY*; *trnM-trnH-trnQ-trnG-trnS1-trnI-trnF-trnA-trnL1-trnE-trnC*). Whilst the *P. rubra* mitochondrial sequence has a large degree of overlap between adjacent genes, the opposite is true for *I. pulchra* and *A. ylvae*. Unlike other metazoan mitochondrial genomes, where genes are adjacent or overlapping and one or two larger non-coding regions are commonly found, *I. pulchra* non-coding sequence is found consistently between protein-coding genes and

Feature	Strand	Start	Stop	Length (nucleotides)	Length (AA)	Start Codon	Stop Codon	Intergenic
cox1	+	1	1539	1539	513	ATA	TAA	309
nad4l	−	1849	2133	285	95	ATG	TAA	126
nad6	−	2260	2727	468	156	ATG	TAG	460
nad4	−	3188	4531	1344	448	ATA	TAA	424
trnT (tgt)	−	4956	5025	70				498
cox2	−	5524	6171	648	216	ATA	TAA	203
nad3	−	6375	6743	369	123	ATG	TAA	70
trnV (tac)	−	6814	6882	69				4
trnY (gta)	−	6887	6953	67				0
trnP (tgg)	+	6954	7022	69				9
nad5	−	7032	8687	1656	552	ATG	TAA	46
cox3	+	8734	9519	786	262	ATG	TAA	22
trnK (ttt)	+	9542	9611	70				17
cob	−	9629	10714	1086	362	ATT	TAA	56
trnW(tca)	−	10771	10837	67				3
trnR (tcg)	−	10841	10908	68				4
trnN (gtt)	−	10913	10984	72				4
nad2	+	10989	11990	1002	334	ATG	TAA	12
nad1	−	12003	12872	870	290	ATG	TAA	19
trnD (gtc)	−	12892	12954	63				3
trnL2 (taa)	+	12958	13025	68				4
atp6	−	13030	13731	702	234	ATA	TAA	222
trnS2 (tga)	−	13954	14020	67				260
rrnL	+	14281	15089	809				−13
rrnS	+	15077	15814	738				3
trnM (cat)	+	15818	15879	62				−4
trnH (gtg)	+	15876	15945	70				2
trnQ (ttg)	−	15948	16007	60				6
trnG (tcc)	+	16014	16079	66				8
trnS1 (tct)	+	16088	16151	64				10
trnI (gat)	+	16162	16231	70				−1
trnF (gaa)	−	16231	16296	66				10
trnA (tgc)	+	16307	16373	67				10
trnL1 (tag)	−	16384	16450	67				6
trnE (ttc)	−	16457	16522	66				6
trnC (gca)	+	16529	16601	73				18

Table 3. Organisation of the Archaphanostoma ylvae 16.6 kb mitochondrial genome.

between tRNAs, ranging in length from eleven to 277 base pairs. In addition, three long non-coding regions of 788, 1143 and 717 base pairs are found at the start of our sequence; between *trnL1* and *trnN*; and *trnN* and *rrnL*. The A + T content of these three sections are 68.78%, 65.79% and 76.15% respectively. The compositional difference between the 717 base pair non-coding region and the rest of the genome is statistically different ($\chi^2 = 25.629$, $p < 0.0001$), with a higher A + T content indicating that it could function as a transcriptional control region. The A + T content of the 788 base pair non-coding region is not significantly higher than the rest of the sequence ($\chi^2 = 0.85$). Similarly, there is a large portion of intergenic, non-coding sequence in the *A. ylvae* genome. Eight regions of non-coding sequence greater than 100 base pairs are found throughout the genome, with 24 additional smaller intergenic regions, ranging in size from 3 to 70 base pairs. Of the larger non-coding sequences, three have an A + T content that is statistically higher than the entire sequence: 309 nucleotides between *cox1* and *nad4l* ($\chi^2 = 3.944$, $p < 0.1$); 126 nucleotides between *nad4l* and *nad6* ($\chi^2 = 6.964$, $p < 0.01$) and 260 nucleotides between *trnS2* and *rrnL* ($\chi^2 = 9.654$, $p < 0.01$). The *P. rubra* sequence has just one longer non-coding intergenic sequence, of 196 base pairs.

As was already indicated by the 9.7 kb published partial genome[26], the gene arrangement we found in *P. rubra* is unique amongst published metazoan mitochondrial genomes. Similarly, neither *I. pulchra* nor *A. ylvae* show any similarity to other published metazoan mitochondrial genomes (Fig. 6). The species analysed in this study share only the small 'block' of *nad3-atp6-nad4-cob* (*I. pulchra*) and *cob-nad4-nad3* (*P. rubra*). However, the order is reversed between the two, and the genes are distributed across both strands in *I. pulchra*, so it is unlikely that this represents a feature inherited from a common ancestor. To quantify the number of common gene arrangements between the species in this study and other mitochondrial genomes, protein-coding gene and ribosomal

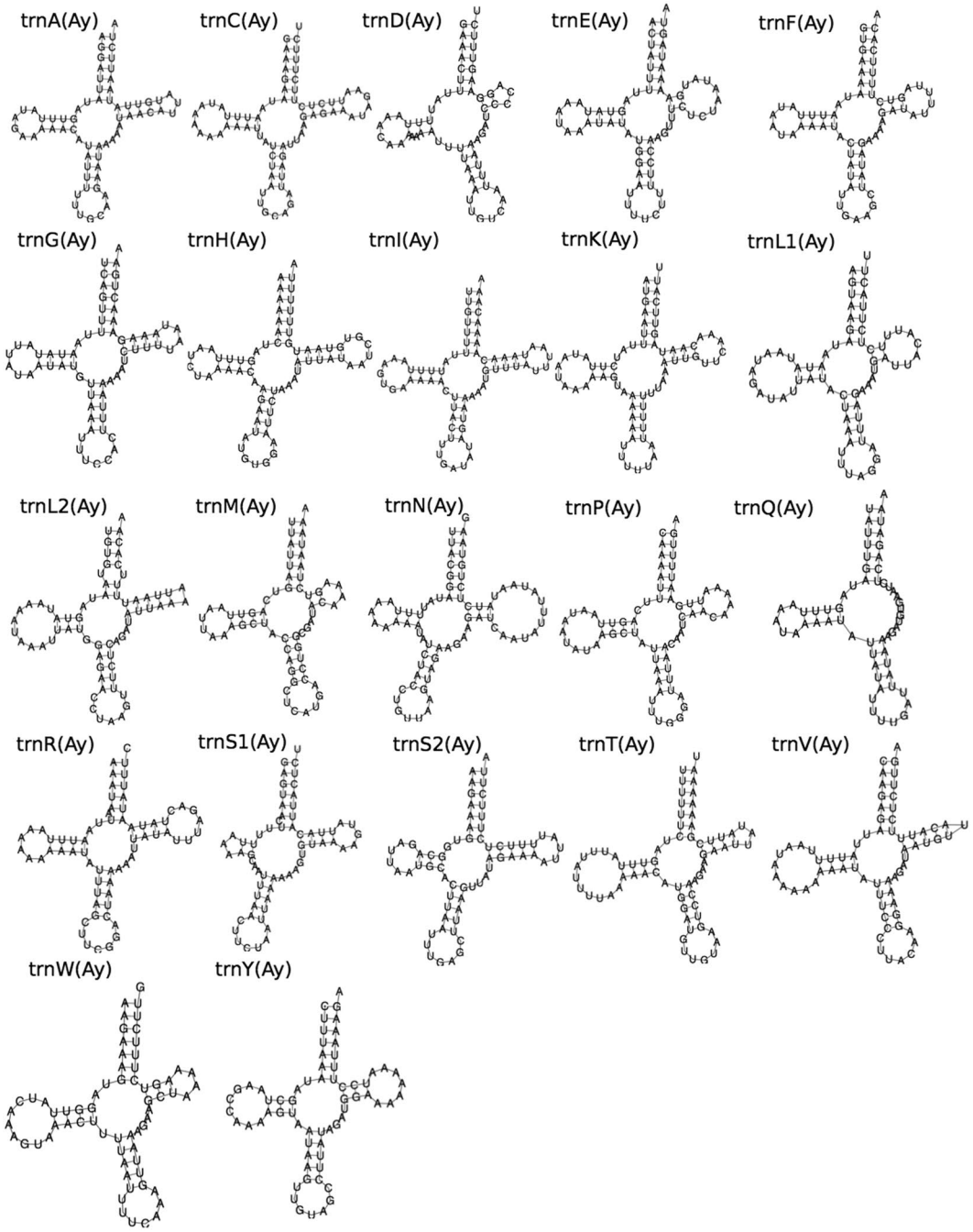


Figure 5. Predicted secondary structure of tRNAs from the mitochondrial genome sequence of *Archaphanostoma ylvae* as predicted by MiTFi in Mitos.

RNA gene order was analysed using CREx[29] (compared to the acoel *S. roscoffensis*, the xenoturbellid *Xenoturbella bocki*, and the metazoan mitochondrial 'ground plan', represented by *L. polyphemus*). Conserved mitochondrial gene 'blocks' (that is, a series of genes, regardless of their order within the grouping) were very infrequent between the species. Of the genomes compared, the highest number of common gene blocks was found between *X. bocki* and *P. rubra*. This result was not significant, finding only 16 common intervals out of a possible 150, and confirming the visual observation that gene order between these species is highly variable.

Phylogenetic analysis, and population differentiation. We used our new mitochondrial data from *P. rubra*, *I. pulchra*, and *A. ylvae* to investigate the internal phylogeny of the acoels and to test support for an Acoela-Xenoturbellida affinity. We observed that including the fast-evolving tunicates into our phylogeny leads to a clustering of these species and the acoels in an artificial long-branched clade (Supplementary Figure S1). Removing the tunicates, Bayesian phylogenetic inference was carried out using the protein-coding genes of *P. rubra*, *I. pulchra* and *A. ylvae* on a trimmed concatenated amino acid alignment, including the additional species

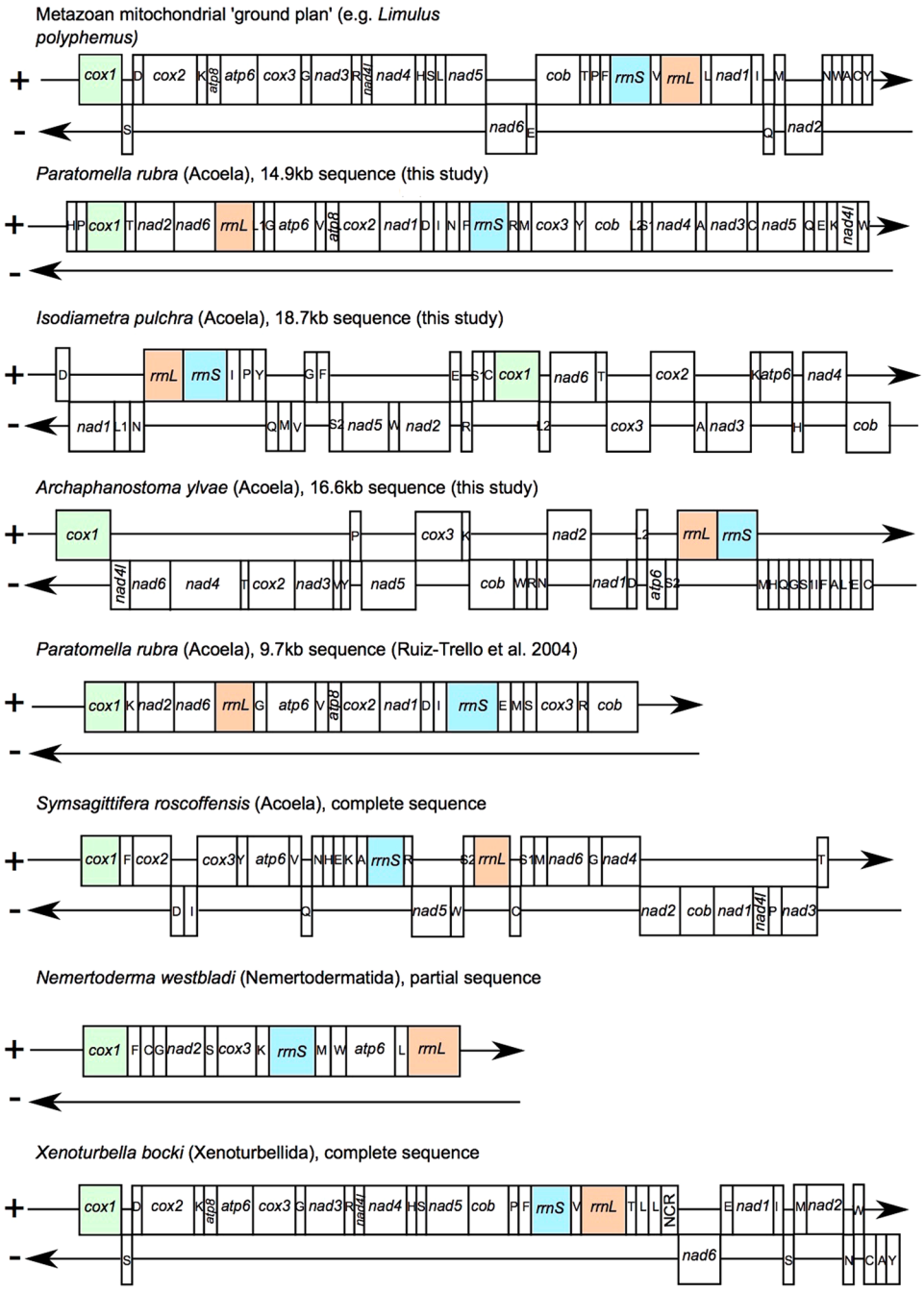


Figure 6. Comparisons of gene orders in the mitochondrial genome sequences resolved for *Paratomella rubra*, *Isodiametra pulchra* and *Archaphanostoma ylvae* compared to a published *P. rubra* fragment; the acoel *Symsagittifera roscoffensis*; the xenoturbellid *Xenoturbella bocki*; the nemertodermatid *Nemertoderma westbladi* and the metazoan mitochondrial 'ground plan' gene order, represented by *Limulus polyphemus*. Genes are not drawn to scale. Coloured genes chosen to show 'anchors' and divergence from the ground plan gene order in other species.

listed in Supplementary Table S2. The data set reached a MaxDiff of 0.17 after 39,525 trees were sampled across 10 chains discarding the first 400 trees (per chain) as burnin and sampling every 10th tree. In both, this and the maximum likelihood approaches, the protostome/deuterostome split was correctly inferred and Xenacoelomorpha were found splitting off inside Deuterostomia. *P. rubra*, *I. pulchra*, and *A. ylvae* were grouped inside Acoela, as expected (Fig. 7).

We detected the Nemertodermatida species *Nemertoderma westbladi* inside Mollusca and we therefore conclude that the limited data available on Genbank for this species is a contamination.

Having access to the published partial *P. rubra* mitochondrial sequence from a population sampled near Barcelona (Spain) and our own samples from Yorkshire, UK, we could estimate total sequence divergence and

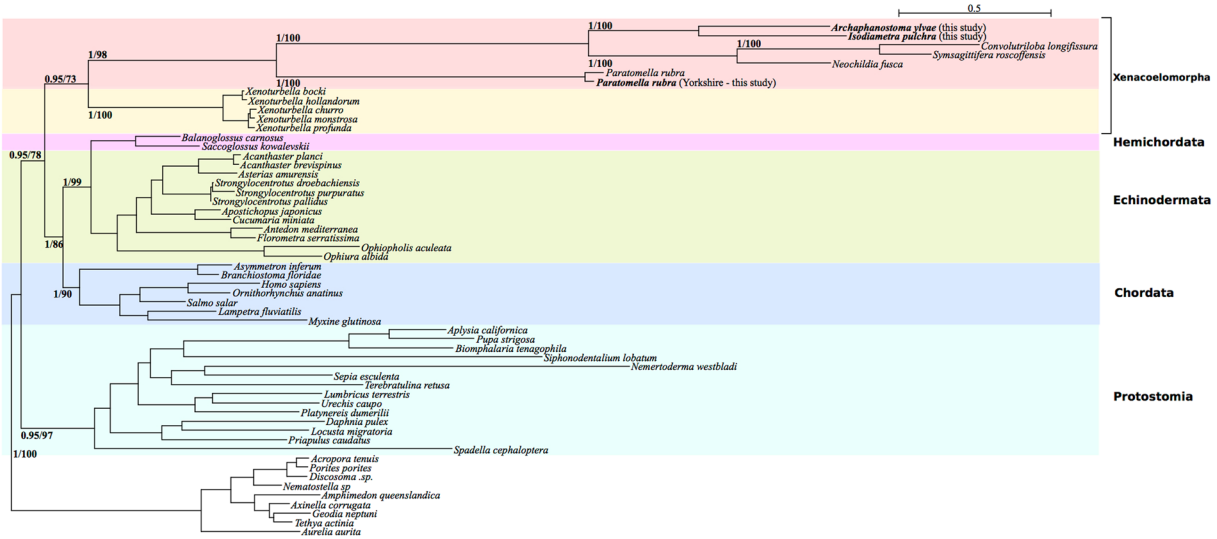


Figure 7. Bayesian (using PhyloBayes[53]) and Maximum Likelihood (using RAxML[51]) phylogenetic analysis of mitochondrial protein-coding genes from the Metazoa, including *P. rubra*, *I. pulchra* and *A. ylvae* with posterior probability and bootstrap support values, respectively, at relevant nodes. Analysis carried out on trimmed alignment. Topology of both trees is identical.

compare non-synonymous to synonymous substitutions in eight protein-coding genes found on the Spanish fragment to the same genes from the Yorkshire mitochondrial genome. We found the 9.7 kb sequences to be 82.62% similar at the nucleotide level. The number of substitutions varied between, for example, 23 in the shortest gene alignment (*atp8*; 177 bp), to 161 in *nad2* (972 bp), and 116 in the 1401 bp long *cox1* alignment (Supplementary Table S3). Notably, non-synonymous substitutions appear to be frequent with, for example, 13 in *atp8*, 104 in *nad2*, and 25 in *cox1* (see Supplementary Table S3). Similarity of the *cox1* sequences on the nucleotide level is only 91% over 666 bp, thus higher when compared to species pairs (Supplementary Table S4), but lower than the 95–98% percent threshold used to distinguish species in *cox1* based barcoding[30].

Discussion

The 14.9 kb sequence of the *P. rubra* mitochondrial genome determined in our analysis contains the full complement of 37 genes typical of metazoan mitochondrial DNA. Numerous lab-based and computational efforts to close the circular mitochondrial genome were unsuccessful, but the complete gene complement and length of our final sequence indicates that this fragment covers the majority of the *P. rubra* complete mitochondrial genome. The difficulty we encountered in attempting to close the circular mitochondrial sequence may be attributed to the AT-rich, repetitive sequence found at both ends of the fragment, which could have prevented PCR amplification. Similar regions have been shown as problematic in studies of other mitochondrial genomes[31]. As no long stretch of non-coding sequence was found for this species in our study, the missing sequence might represent its mitochondrial transcription control region. Nonetheless, the overall AT content of the *P. rubra* mitochondrial sequence (78.15%) is high even for mtDNA, and greater than the A + T content of the mitochondrial genome of the acoel *S. roscoffensis* (75.3%)[11] and the published partial *P. rubra* genome (76.4%)[26].

The validity of the duplicated sequence found in our analysis of the *I. pulchra* mitochondrial genome could not be confirmed by PCR or computational efforts to map short reads to resolve it. Duplications within mitochondrial genomes are not uncommon, and changes to mitochondrial gene order are widely thought to arise as a result of a sequence 'duplication and deletion' mechanism[14,32,33]. A number of mitochondrial genomes with duplicated sequences have been reported in species with a divergent mitochondrial gene order[33–35]. Given the highly unusual gene order of the *I. pulchra* mitochondrial genome, a genomic duplication such as this could provide evidence for a genomic 'duplication and deletion' rearrangement of genes. The rearrangement and separation of protein-coding genes in other mitochondrial genomes has been attributed to long, non-tandem, inverted repeats[34], and this could be true for *I. pulchra*. Furthermore, very long nematode mitochondrial genomes with variable duplicated regions have been found with a conserved region containing the majority of the protein-coding genes[36]: in *I. pulchra* the protein-coding genes and tRNAs, with the exception of *nad1* and *trnD* and *L1* are found in one long region, outside of the duplicated section. However, long non-coding duplications are frequently adjacent to tRNA*s* or other sequences capable of forming stem-and-loop structures[37]. This is not true for the potential duplicate in *I. pulchra*. Most puzzling, both occurrences of the duplicate are identical, nucleotide-by-nucleotide, and unless the duplication occurred exceptionally recently, it is likely that spontaneous mutations would result in differences between the two copies of the sequence, especially given the elevated mutation rate of mitochondrial genomes. While it is true that the duplicated sequences appear at the start and end point of transcriptome assembly contigs, meaning it is possible that the duplication observed occurred only as a result of a sequencing and assembly error, their existence is nevertheless supported by PCR products which show an identical sequence being adjacent to both *rrnL* and to *cob*.

	Isodiametra pulchra	*Paratomella rubra*	*Symsagittifera roscoffensis*	*Archaphanostoma ylvae*
cox1	1536	1563	1551	1539
cox2	615	663	741	648
cox3	798	786	792	786
nad1	881	1053	870	870
nad2	1053	1014	990	1002
nad3	378	390	393	369
nad4	1344	1326	1350	1344
nad4l	absent	309	270	285
nad5	1710	1752	1776	1656
nad6	476	462	480	468
cob	1134	1083	1161	1086
atp6	681	609	702	702
atp8	absent	177	absent	absent

Table 4. Length of protein–coding genes in acoel mitochondrial genomes. All gene lengths in base pairs.

The 14.9 kb mitochondrial genome of *P. rubra*, the 18.7 kb sequence from *I. pulchra* and the complete 16.6 kb *A. ylvae* mitochondrial genome show no significant organisational similarity to any other published metazoan mitochondrial genome (Fig. 6). Comparison of the 14.9 kb *P. rubra* sequence with the published 9.7 kb *P. rubra* fragment shows an identical protein-coding and ribosomal gene order, but with variation in tRNA order (Fig. 6). tRNAs are reported to show much more frequent gene translocation compared to larger genes[38], which could account for these discrepancies. This, and the differences on the nucleotide level, including the relatively low level of similarity in the *cox1* barcoding gene might indicate that *P. rubra* collected from Barcelona (Spain)[26] and our animals from Yorkshire (England) should be regarded as cryptic species and not just divergent populations. Given the large and mostly unresolved diversity in benthic communities[39], and the marine environment in general[40], a differentiation into (cryptic) species cannot be seen as surprising.

All species analysed in this study are unique in the orientations and orders of their genes: *P. rubra* has genes exclusively on one strand; *I. pulchra* has an almost-equal distribution of genes across both the plus (18 genes) and minus (17 genes) strand; with *cox1* in a 'forward' orientation at the start of the genome, the majority of the protein-coding genes for *A. ylvae* are found on the minus strand. Furthermore, genes in *I. pulchra* and *A. ylvae* are not clustered into groups of 'gene blocks' on the same strand, but are found frequently as one or two genes on each strand. The finding of a unique gene order for these species seems to be typical for the acoels: analysis of the complete *S. roscoffensis* mitochondrial genome found no gene order similarity to any other species published to date, suggesting great variability in mitochondrial gene order amongst the acoels[11].

In addition to an unusual gene order, the mitochondrial genome of *P. rubra* shows frequent overlaps between protein-coding genes and tRNAs. tRNAs have been reported within protein-coding genes in other metazoan mitochondrial genomes[41,42], and given that no other location could be predicted for these sequences, this overlap could represent the simultaneous coding for both tRNAs and protein-coding genes. Overlap in coding sequence could be the result of selection to reduce genome size, accompanied by a reduction in non-coding sequence[42], and truncated tRNAs with incomplete secondary structure[43], both of which are also found for the *P. rubra* sequence. The opposite is true for the *I. pulchra* and *A. ylvae* sequences. For *I. pulchra*, the sequence we could confidently verify makes the minimal length of the *I. pulchra* mitochondrial genome 18,725 base pairs, and it is likely to be longer in the complete closed circular genome. As has been found for other 'long' mitochondrial genomes, this is largely due to a large portion of the genome being non-coding[44]. The lengths of protein-coding genes inferred for *I. pulchra* are similar to those of other acoel species (Table 4), and in addition, two protein-coding genes (*atp8* and *nad4l*) appear to have been lost from the genome, contributing to a reduced proportion of protein-coding gene sequence within the genome. The loss of *atp8* is not unusual, and has been reported in a number of unrelated taxa, as well as *S. roscoffensis* and in our *A. ylvae* mitochondrial genome[11]. The absence of *nad4l* is more unusual, and could be a result of its existence in a portion of the genome that we have been unable to sequence. Although non-coding sequence contributes a relatively large proportion of the *A. ylvae* mitochondrial genome (17.17% compared to 22.72% in the *I. pulchra* sequence), the total genome is not exceptionally long.

The internal phylogeny we resolve for Acoela is in line with that proposed by Jondelius *et al.*[25]. *I. pulchra* and *A. ylvae* group together in the family Isodiametridae. Isodiametridae groups with *S. roscoffensis*, *Neochildia fusca* and *Convolutriloba longifissura*, which are all members of the Convolutidae. *P. rubra* forms a separate branch outside the Convolutidae, representing the Paratomellidae. We interpreted the initial grouping of the acoels and tunicates as a classical example of long branch attraction (LBA) (Supplementary Figure S1). The accelerated substitution rates in mitochondrial DNA are also evidenced by the cryptic divergence we find in *P. rubra*, and may well lead to LBA in phylogenies derived from mitochondrial protein-coding genes, owing to the clustering of rapidly evolving lineages. This is of particular relevance for acoel species, which already demonstrate a very rapid rate of nucleotide substitution compared to other metazoans, leaving them vulnerable to LBA. Excluding the urochordates, we do retrieve Xenacoelomorpha as a branch of the deuterostomes, as expected (Fig. 7).

The mitochondrial genome sequences we analysed for the acoel species *P. rubra*, *I. pulchra* and *A. ylvae* have very divergent gene orders compared to other metazoan species. Furthermore, these species have very different

mitochondrial features: a large amount of genomic overlap in *P. rubra*, and a lot of non-coding sequence in *I. pulchra* and *A. ylvae*. It is also possible that the mitochondrial genome of *I. pulchra* has a non-tandem inverted duplication - which could provide a mechanism for gene order variation - but this could not be confirmed by lab or computational based methods. Although limited to four species, the uniqueness of acoel mitochondrial genomes analysed so far[11,26] means that gene order and other mitochondrial genome features may not be phylogenetically informative for this order, although further mitochondrial genomes from other members of the Acoela would no doubt aid in this comparative analysis. Similarly, the cryptic divergence found between *P. rubra* samples from Yorkshire and Barcelona illustrate the usefulness of studying mitochondrial genomes to understand hidden species diversity. Our data clearly emphasise the still problematic placing of Xenacoelomorpha, with the clade firmly placed inside deuterostomes, but LBA drawing the acoels towards the outgroup when the tunicates were included. In summary, more data from genomes of early branching taxa are needed to resolve phylogenetic and biological questions.

Methods

Specimen collection, DNA extraction and PCR. Live *Paratomella rubra* specimens were isolated from sand samples collected from Filey, North Yorkshire and were immediately frozen and stored at −70 °C following identification. Specimens of *Isodiametra pulchra* were cultured in petri dishes with nutrient-enriched f/2 sea water and fed *ad libitum* on *Nitzschia curvilineata* diatoms. DNA was extracted from live specimens of *I. pulchra* and frozen specimens of *P. rubra* using the QIAamp DNA Micro Kit (Qiagen: Product No. 56304) with the manufacturer recommended protocol.

All PCRs were done using the GeneAmp PCR System 2700 (Applied Bioscience). PCRs were carried out using the Expand Long-Range PCR Kit (Roche Applied Sciences: Product No. 11681834001), following manufacturer recommendations for 50 μl reaction set-up. General cycling protocol was: 92 °C for 2 min; 15 cycles of: 92 °C for 10 sec, 57 °C for 15 sec, 68 °C at initial elongation time (approximated as 1 min per 1000 base pairs to be amplified); 2 cycles each of: 92 °C for 10 sec, 57 °C for 15 sec, 68 °C at 40 sec longer than initial elongation time, repeated at increasing 40 sec intervals for a further 14 cycles; a final elongation stage at 68 °C for 7 min and a 4 °C 'hold' stage. Where PCRs were not successful using this protocol, they were repeated using the Q5 High-Fidelity PCR Kit (New England Biolabs: Product No. E0555L), following manufacturer recommendations for a 25 μl reaction. Cycling protocol was: 92 °C for 2 min; 40 cycles of: 92 °C for 10 sec, 58 °C for 15 sec, 68 °C at initial elongation time (approximated as 1 min per 1000 base pairs to be amplified); a final elongation stage at 68 °C for 7 min. Amplified products were visualised on ethidium-bromide stained TAE 0.8% gels. Bands of expected size were purified using the High Pure PCR Product Purification Kit (Roche Applied Sciences: Product No. 11732668001) and sent for sequencing by Source BioScience Life Sciences. Only amplifications that resulted in one clear band on the TAE 0.8% agarose gel were sequenced.

Three fragments of sequence from the mitochondrial genome of *P. rubra*, of size ~5.8 kb, ~4 kb, and ~1.2 kb, were generated from our gDNA assembly. Fragments were verified using a translated nucleotide query blast with invertebrate codon usage (blastx NCBI), and their orientation determined by gene annotation in comparison to the published 9.7 kb section of the *P. rubra* mitochondrial genome[26]. Primers were designed in conserved gene regions to:

(1) Amplify across the 'N-stretches' present in the 5.8 kb and 4 kb fragments (8 and 9 N-stretches respectively, all of arbitrary length 50 base pairs).

(2) Cover the whole 1.2 kb fragment, with the aim of resolving the two frameshift mutations within the assembled sequence.

(3) Close the circular mitochondrial genome, by joining the 5.8 kb fragment to the 1.2 kb fragment; the 1.2 kb fragment to the 4 kb fragment; and the 4 kb fragment to the 5.8 kb fragment (see Supplementary Figure S5). Amplification of the fragments joining the 1.2 kb fragment to the 4 kb fragment and to close the mitochondrial genome using standard PCR cycling were unsuccessful. These were repeated using a touchdown protocol with Expand Long-Range polymerase. Annealing temperature was set at 65 °C with decreasing 2 °C intervals every 2 cycles down to 49 °C. Initial elongation time was calculated as before, increasing 30 sec every two cycles of the touchdown, with a final 6 cycles at 49 °C. This successfully amplified the region joining the 1.2 kb fragment to the 4 kb fragment, but we could not close the circular genome. Design of three new forward and reverse primers, tried in all combinations and using variable PCR parameters were unsuccessful in closing the mitochondrial genome. Additional RNA-Seq and DNA genomic sequencing data corroborated the stretches of sequence at either end of the mitochondrial genome but did not aid in closing the circle.

Three mitochondrial contigs of size ~13 kb, ~3.5 kb and ~1.3 kb were identified from *I. pulchra* Trinity transcriptome assembly from total RNA sequencing. A further contig of ~19 kb was also identified, covering the entire ~1.3 and 13 kb regions, and ~2.4 kb of the 3.5 kb sequence. Fragments were verified using blastx, NCBI, as outlined for *P. rubra*, and approximations for the location of protein-coding genes and tRNAs determined using the MITOS mitochondrial genome annotation server[45] (http://mitos.bioinf.uni-leipzig.de/help.py). Primers were designed to span the 13 kb contig in two ~5 kb sections, and to join the 13 kb contig to the 3.5 kb contig in both directions, to close the mitochondrial genome and check the validity of the duplicated region (Fig. 3). RNA-Seq data for *I. pulchra* were mapped to the long transcriptome assembly contigs and PCR sequencing results using NextGenMap[46], and visualised using Tablet[47].

We accidentally co-sequenced *A. ylvae* at very low coverage as part of a *P. rubra* genome sequencing experiment. From an initial paired end assembly of Illumina HiSeq data with the CLC assembly cell software (v.5.0) we extracted the full mitochondrial circle of *A. ylvae* in a single contig identified with BLAST. This was then annotated using MITOS and manual refinement as described above.

Genome annotation. For *P. rubra*, all sequenced fragments were aligned against the initial scaffold of the 9.7 kb published sequence[26]; the 5.8 kb, 4 kb and 1.2 kb genome assembly sequences; and an additional long genome assembly fragment of length 14,954 (see Supplementary Figure S5). All contigs and PCR sequencing results were similarly aligned for *I. pulchra*, but without a reference sequence (see Fig. 3). Alignments were visualised using Mesquite (http://mesquiteproject.org) with invertebrate mitochondrial translated amino acid state colour coding. Where ambiguity remained between PCR sequencing results and genome or transcriptome assembly fragments, the genome or transcriptome assembly nucleotide sequence was used to establish a final 'consensus' sequence for each mitochondrial genome. This was of particular relevance for repetitive AT regions – for example, within *P.rubra nad1*, for which PCR sequencing results were inconclusive. In the case of *I. pulchra*, where the validity of the duplicated sections could not be confidently determined, we resolved a consensus sequence of 18, 725 base pairs (Fig. 3).

The region for each protein-coding mitochondrial gene (*nad1–6, nad4l, cox1–3, cob, atp6 and atp8*) in the *P. rubra*, *I. pulchra* and *A. ylvae* sequences were compared against published mitochondrial genomes using a translated nucleotide query (blastx, NCBI) with NCBI translation table number 5 'invertebrate mitochondrial'. Published genes from the mitochondrial genomes of the acoels *Symsagittifera roscoffensis* and *P.rubra* were downloaded from the NCBI Nucleotide database and aligned to the new consensus gene sequences of both *P. rubra*, *I. pulchra* and *A. ylvae* to verify the location of protein-coding and ribosomal RNA-encoding genes. The 5′ end of protein-coding genes were inferred to start from the first in-frame start codon (ATN, GTG, TTG, or GTT), even if this appeared to overlap with the preceding gene. Similarly, the terminal of protein-coding genes was inferred to be the first in-frame stop codon (TAA, TAG, or TGA). If no stop codon was present, a truncated stop-codon (T– or TA-) prior to the beginning of the next gene was assumed to be the termination codon, completed by post-transcription polyadenlylation. tRNA sequences and putative secondary structures were identified using the Mitfi program within MITOS.

Sequence alignment, phylogenetic analysis, and evolutionary rates. Phylogenetic analysis was performed using a concatenated amino acid alignment of all thirteen protein-coding genes for *P. rubra*, and all eleven protein-coding genes present in *I. pulchra*. Nucleotide sequences from all three acoel taxa and an additional set of species comprising 54 metazoans, taken from a range of published metazoan mitochondrial genomes representing deuterostomes, protostomes, cnidarians, and two species of poriferans as an outgroup (Supplementary Table S2) were aligned using TranslatorX (http://www.translatorx.co.uk/) independently for all genes with the appropriate mitochondrial genetic code set for each taxon included, using ClustalOmega[48] for amino acid alignment (Supplementary Table S2). Protein alignments were reduced to the most informative residues using trimAl v.1.4.rev15[49] with standard settings. Regions showing ambiguity in alignment were excluded, so that only blocks of well-aligned sequence were included for analysis.

We initially re-constructed neighbour nets in SplitsTrees v.4[50] to screen our dataset for potentially non-treelike patterns, which could impede our phylogenetic analysis. Subsequently, we used RAxML[51] v. 8.2.9 to infer maximum likelihood phylogenies from the original and the reduced alignments under the MTZOA model[52]. Bootstrapping was conducted employing the 'autoMRE' option in RAxML and the trees visualised with figtree v.1.4 (http://tree.bio.ed.ac.uk/software/figtree). We carried out Bayesian inference on the same trimmed alignment with PhyloBayes v.4.1[53] under the MTZOA model. We ran 10 chains in parallel and stopped the tree search at ~3950 trees per chain, with a maximum difference of 0.17, when discarding 400 trees as burnin and sampling every 10th tree per chain.

We used the Geneious software (v.8) to calculate sequence differences between a 9.7 kb section of the *P. rubra* mitochondrial genome originating from worms sampled in Filey, Yorkshire (UK) and Barcelona (Spain)[26]. For eight protein-coding genes found on this section we used ParaAT (v2.0)[54] to calculate translation alignments and the KaKs calculator (v1.2)[55] to access substitution rates. Also using Geneious we calculated a difference matrix for the *cox1* barcoding gene of the two *P. rubra* populations in comparison to acoel coI sequences retrieved from Genbank.

References

1. Bourlat, S. J. & Hejnol, A. Acoels. *Curr Biol* **19**, R279–280, doi:10.1016/j.cub.2009.02.045 (2009).
2. Ruiz-Trillo, I., Riutort, M., Littlewood, D. T., Herniou, E. A. & Baguña, J. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* **283**, 1919–1923, doi:10.1126/science.283.5409.1919 (1999).
3. Katayama, T., Yamamoto, M., Wada, H. & Satoh, N. Phylogenetic position of Acoel turbellarians inferred from partial 18S rDNA sequences. *Zoolog Sci* **10**, 529–536 (1993).
4. Philippe, H., Brinkmann, H., Martinez, P., Riutort, M. & Baguñà, J. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. *PLoS One* **2**, e717, doi:10.1371/journal.pone.0000717 (2007).
5. Lundin, K. The epidermal ciliary rootlets of *Xenoturbella bocki* (Xenoturbellida) revisited: new support for a possible kinship with the Acoelomorpha (Platyhelminthes). *Zoologica Scripta* **27**, 8–270, doi:10.1111/zsc.1998.27.issue-3 (1998).
6. Raikova, O. I., Reuter, M., Jondelius, U. & Gustafsson, M. K. The brain of the Nemertodermatida (Platyhelminthes) as revealed by anti-5HT and anti-FMRFamide immunostainings. *Tissue Cell* **32**, 358–365, doi:10.1054/tice.2000.0121 (2000).
7. Hejnol, A. *et al.* Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci* **276**, 4261–4270, doi:10.1098/rspb.2009.0896 (2009).
8. Cannon, J. T. *et al.* Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530**, 89–93, doi:10.1038/nature16520 (2016).
9. Philippe, H. *et al.* Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature* **470**, 255–258, doi:10.1038/nature09676 (2011).
10. Boore, J. L. Animal mitochondrial genomes. *Nucleic Acids Res* **27**, 1767–1780, doi:10.1093/nar/27.8.1767 (1999).
11. Mwinyi, A. *et al.* The phylogenetic position of Acoela as revealed by the complete mitochondrial genome of Symsagittifera roscoffensis. *BMC Evol Biol* **10**, 309, doi:10.1186/1471-2148-10-309 (2010).
12. Telford, M. J., Herniou, E. A., Russell, R. B. & Littlewood, D. T. Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proc Natl Acad Sci USA* **97**, 11359–11364, doi:10.1073/pnas.97.21.11359 (2000).

13. Saccone, C., De Giorgi, C., Gissi, C., Pesole, G. & Reyes, A. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* **238**, 195–209, doi:10.1016/S0378-1119(99)00270-X (1999).
14. Boore, J. L. & Brown, W. M. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr Opin Genet Dev* **8**, 668–674, doi:10.1016/S0959-437X(98)80035-X (1998).
15. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of Xenoturbella and the position of Xenacoelomorpha. *Nature* **530**, 94–97, doi:10.1038/nature16545 (2016).
16. Perseke, M. *et al*. The mitochondrial DNA of Xenoturbella bocki: genomic architecture and phylogenetic analysis. *Theory in Biosciences* **126**, 10.1007/s12064-007-0007-7 (2007).
17. Bourlat, S. J., Rota-Stabelli, O., Lanfear, R. & Telford, M. J. The mitochondrial genome structure of Xenoturbella bocki (phylum Xenoturbellida) is ancestral within the deuterostomes. *BMC Evolutionary Biology* **9**, 107, doi:10.1186/1471-2148-9-107 (2009).
18. Moritz, C. & Brown, W. M. Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. *Proc Natl Acad Sci USA* **84**, 7183–7187, doi:10.1073/pnas.84.20.7183 (1987).
19. Boore, J. L., Lavrov, D. V. & Brown, W. M. Gene translocation links insects and crustaceans. *Nature* **392**, 667–668, doi:10.1038/33577 (1998).
20. Crezee, M. Paratomella rubra, Rieger and Ott, an amphiatlantic acoel turbellarian. *Cahiers De Biologie Marine* **19**, 1–9 (1978).
21. Rieger, R. & Ott, J. Gezeitenbedingte Wanderungen von Turbellarien und Nematoden eines nordadriatischen Sandstrandes. *Vie Milieu (Suppl.)* **22**, 425–447 (1971).
22. Achatz, J. G. & Martinez, P. The nervous system of Isodiametra pulchra (Acoela) with a discussion on the neuroanatomy of the Xenacoelomorpha and its evolutionary implications. *Front Zool* **9**, 27, doi:10.1186/1742-9994-9-27 (2012).
23. Kånneby, T., Bernvi, D. C. & Jondelius, U. Distribution, delimitation and description of species of Archaphanostoma (Acoela). *Zoologica Scripta* **44**, 218–231, doi:10.1111/zsc.12092 (2015).
24. Achatz, J. G., Chiodin, M., Salvenmoser, W., Tyler, S. & Martinez, P. The Acoela: on their kind and kinships, especially with nemertodermatids and xenoturbellids (Bilateria incertae sedis). *Organisms, Diversity & Evolution* **13**, 267–286, doi:10.1007/s13127-012-0112-4 (2013).
25. Jondelius, U., Wallberg, A., Hooge, M. & Raikova, O. I. How the Worm Got its Pharynx: Phylogeny, Classification and Bayesian Assessment of Character Evolution in Acoela. *Systematic Biology* **60**, 845–871, doi:10.1093/sysbio/syr073 (2011).
26. Ruiz-Trillo, I., Riutort, M., Fourcade, H. M., Baguna, J. & Boore, J. L. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol Phylogenet Evol* **33**, 10.1016/j.ympev.2004.06.002 (2004).
27. De Mulder, K. *et al*. Characterization of the stem cell system of the acoel Isodiametra pulchra. *BMC Developmental Biology* **9**, 1–17, doi:10.1186/1471-213x-9-69 (2009).
28. Egger, B. *et al*. To be or not to be a flatworm: the acoel controversy. *Plos One* **4** (2009).
29. Bernt, M. *et al*. CREx: inferring genomic rearrangements based on common intervals. *Bioinformatics* **23**, 2957–2958, doi:10.1093/bioinformatics/btm468 (2007).
30. Hebert, P. D. N., Ratnasingham, S. & de Waard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London. *Series B: Biological Sciences* **270**, S96–S99, doi:10.1098/rsbl.2003.0025 (2003).
31. Sakai, M. & Sakaizumi, M. The complete mitochondrial genome of Dugesia japonica (Platyhelminthes; order Tricladida). *Zoolog Sci* **29**, 672–680, doi:10.2108/zsj.29.672 (2012).
32. Moritz, C., Dowling, T. E. & Brown, W. M. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annu Rev Ecol Syst* **18**, 10.1146/annurev.es.18.110187.001413 (1987).
33. Braband, A., Podsiadlowski, L., Cameron, S. L., Daniels, S. & Mayer, G. Extensive duplication events account for multiple control regions and pseudo-genes in the mitochondrial genome of the velvet worm Metaperipatus inae (Onychophora, Peripatopsidae). *Molecular Phylogenetics and Evolution* **57**, 293–300, doi:10.1016/j.ympev.2010.05.012 (2010).
34. Hyman, B. C., Beck, J. L. & Weiss, K. C. Sequence amplification and gene rearrangement in parasitic nematode mitochondrial DNA. *Genetics* **120**, 707–712 (1988).
35. Zhou, X., Lin, Q., Fang, W. & Chen, X. The complete mitochondrial genomes of sixteen ardeid birds revealing the evolutionary process of the gene rearrangements. *BMC Genomics* **15**, 573, doi:10.1186/1471-2164-15-573 (2014).
36. Hyman, B. C., Lewis, S. C., Tang, S. & Wu, Z. Rampant gene rearrangement and haplotype hypervariation among nematode mitochondrial genomes. *Genetica* **139**, 611–615, doi:10.1007/s10709-010-9531-3 (2011).
37. Stanton, D. J., Daehler, L. L., Moritz, C. C. & Brown, W. M. Sequences with the potential to form stem-and-loop structures are associated with coding-region duplications in animal mitochondrial DNA. *Genetics* **137**, 233–241 (1994).
38. Bermudez-Santana, C. *et al*. Genomic organization of eukaryotic tRNAs. *BMC Genomics* **11**, 1–14, doi:10.1186/1471-2164-11-270 (2010).
39. Fonseca, V. G. *et al*. Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications* **1**, 98, doi:10.1038/ncomms1095 (2010).
40. Schiffer, P. H. & Herbig, H.-G. Endorsing Darwin: global biogeography of the epipelagic goose barnacles Lepas spp. (Cirripedia, Lepadomorpha) proves cryptic speciation. *Zoological Journal of the Linnean Society* **177**, 507–525, doi:10.1111/zoj.12373 (2016).
41. Morrison, D. A. How and where to look for tRNAs in Metazoan mitochondrial genomes, and what you might find when you get there. *arXiv.org* (2010).
42. He, Y., Jones, J., Armstrong, M., Lamberti, F. & Moens, M. The Mitochondrial Genome of Xiphinema americanum sensu stricto (Nematoda: Enoplea): Considerable Economization in the Length and Structural Features of Encoded Genes. *Journal of Molecular Evolution* **61**, 819–833, doi:10.1007/s00239-005-0102-7 (2005).
43. Doublet, V. *et al*. Large gene overlaps and tRNA processing in the compact mitochondrial genome of the crustacean Armadillidium vulgare. *RNA Biology* **12**, 1159–1168, doi:10.1080/15476286.2015.1090078 (2015).
44. Minxiao, W., Song, S., Chaolun, L. & Xin, S. Distinctive mitochondrial genome of Calanoid copepod Calanus sinicus with multiple large non-coding regions and reshuffled gene order: Useful molecular markers for phylogenetic and population studies. *BMC Genomics* **12**, 73, doi:10.1186/1471-2164-12-73 (2011).
45. Bernt, M. *et al*. MITOS: improved *de novo* metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* **69**, 313–319, doi:10.1016/j.ympev.2012.08.023 (2013).
46. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791, doi:10.1093/bioinformatics/btt468 (2013).
47. Milne, I. *et al*. Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**, 193–202, doi:10.1093/bib/bbs012 (2012).
48. Sievers, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539–539, doi:10.1038/msb.2011.75 (2011).
49. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973, doi:10.1093/bioinformatics/btp348 (2009).
50. Huson, D. H. & Bryant, D. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* **23**, 254–267, doi:10.1093/molbev/msj030 (2006).
51. Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, doi:10.1093/bioinformatics/btu033 (2014).

52. Rota-Stabelli, O., Yang, Z. & Telford, M. J. MtZoa: A general mitochondrial amino acid substitutions model for animal evolutionary studies. *Molecular Phylogenetics and Evolution* (2009).
53. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288, doi:10.1093/bioinformatics/btp368 (2009).
54. Zhang, Z. *et al.* ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochemical and Biophysical Research Communications* **419**, 779–781, doi:10.1016/j.bbrc.2012.02.101 (2012).
55. Zhang, Z. *et al.* KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging. *Genomics, Proteomics & Bioinformatics* **4**, 259–263, doi:10.1016/S1672-0229(07)60007-2 (2006).

Acknowledgements

This work was supported by the European Research Council (ERC-2012-AdG 322790), the Leverhulme Trust (grant F/07 134/DA) and the Biotechnology and Biological Sciences Research Council (grant BBS/H006966/1). MJT was supported by a Royal Society Wolfson Research Merit Award. We would like to thank Richard Copley (The Wellcome Trust Centre for Human Genetics, University of Oxford) and the Oxford Wellcome Trust for their assistance with *P. rubra* sequencing, We are indebted to Anne Zakrzewski for her assistance with animal collection trips, associated animal identification, and valuable comments on the manuscript.

Author Contributions

Conceived and designed the experiments: H.E.R., F.L., B.E. Performed the experiments: H.E.R., B.E., F.L., P.H.S. Analysed the data: H.E.R., P.H.S., M.J.T. Wrote the paper: H.E.R., M.J.T., P.H.S.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-01608-4

Competing Interests: The authors declare that they have no competing interests.

Accession codes: The mitochondrial sequences for *P. rubra*, *A. ylvae* and *I. pulchra* have been submitted to NCBI GenBank under Accession Numbers KY825222, KY825223 and KY825224, respectively.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

PLOS | ONE

RESEARCH ARTICLE

The Complete Mitochondrial Genome of the Geophilomorph Centipede *Strigamia maritima*

Helen E. Robertson[1☯], François Lapraz[1☯¤], Adelaide C. Rhodes[2], Maximilian J. Telford[1]*

1 Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London, United Kingdom, **2** Center for Genome Research and Biocomputing, 2750 SW Campus Way, Oregon State University, Corvallis, Oregon, United States of America

☯ These authors contributed equally to this work.
¤ Current address: CNRS, CBD UMR5547, Université de Toulouse, UPS, CBD (Centre de Biologie du Développement), Bâtiment 4R3, 118 route de Narbonne, Toulouse, France
* m.telford@ucl.ac.uk

OPEN ACCESS

Citation: Robertson HE, Lapraz F, Rhodes AC, Telford MJ (2015) The Complete Mitochondrial Genome of the Geophilomorph Centipede *Strigamia maritima*. PLoS ONE 10(3): e0121369. doi:10.1371/journal.pone.0121369

Academic Editor: Bi-Song Yue, Sichuan University, CHINA

Received: December 12, 2014

Accepted: January 31, 2015

Published: March 20, 2015

Data Availability Statement: Data availability-Strigamia maritima complete mitochondrial genome sequence submitted to NCBI GenBank with Accession Number KP173664.

Funding: H.E.R. is supported by the ERC (ERC-2012-AdG 322790-XENOTURBELLA), F.L. is supported by the Biotechnology and Biological Sciences Research Council (BBS/B/0675X), and M.J.T. is supported by a Royal Society Wolfson Research Merit Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Strigamia maritima (Myriapoda; Chilopoda) is a species from the soil-living order of geophilomorph centipedes. The Geophilomorpha is the most speciose order of centipedes with over a 1000 species described. They are notable for their large number of appendage bearing segments and are being used as a laboratory model to study the embryological process of segmentation within the myriapods. Using a scaffold derived from the recently published genome of *Strigamia maritima* that contained multiple mitochondrial protein-coding genes, here we report the complete mitochondrial genome of *Strigamia*, the first from any geophilomorph centipede. The mitochondrial genome of *S. maritima* is a circular molecule of 14,938 base pairs, within which we could identify the typical mitochondrial genome complement of 13 protein-coding genes and 2 ribosomal RNA genes. Sequences resembling 16 of the 22 transfer RNA genes typical of metazoan mitochondrial genomes could be identified, many of which have clear deviations from the standard 'cloverleaf' secondary structures of tRNA. Phylogenetic trees derived from the concatenated alignment of protein-coding genes of *S. maritima* and >50 other metazoans were unable to resolve the Myriapoda as monophyletic, but did support a monophyletic group of chilopods: *Strigamia* was resolved as the sister group of the scolopendromorph *Scolopocryptos sp*. and these two (Geophilomorpha and Scolopendromorpha), along with the Lithobiomorpha, formed a monophyletic group the Pleurostigmomorpha. Gene order within the *S. maritima* mitochondrial genome is unique compared to any other arthropod or metazoan mitochondrial genome to which it has been compared. The highly unusual organisation of the mitochondrial genome of *Strigamia maritima* is in striking contrast with the conservatively evolving nuclear genome: sampling of more members of this order of centipedes will be required to see whether this unusual organization is typical of the Geophilomorpha or results from a more recent reorganisation in the lineage leading to *Strigamia*.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Strigamia maritima is a geophilomorph centipede found widely along the coasts of North West Europe. It typically inhabits shingle beaches and stone crevices around the high tide line, where it feeds on crustaceans and insect larvae [1]. Geophilomorph centipedes demonstrate a number of unique features that make them a group of particular interest for evolutionary and developmental studies [2–4]. Unlike the vast majority of arthropod species, Geophilomorph members within the clade Adesmata, to which *S. maritima* belongs, show variability in adult segment number within the same species and between sexes [2]. Consequently, they represent an interesting group for studying developmental biology and the evolution of segmentation [2, 5]. Within the geophilomorphs, *S. maritima* is being used as a model species for investigating the evolution of segmentation within the arthropods [2] and understanding developmental processes within the myriapods [3]. A number of studies have been carried out to characterise its embryological development [3, 6–8], and in particular the process of trunk segmentation [5, 9–15].

S. maritima is the first centipede, and indeed the first myriapod, with a completely sequenced nuclear genome [16]. As part of the genome sequencing effort, one of the assembled scaffolds was discovered to contain numerous mitochondrial protein-coding genes, and it was deemed likely that this scaffold represented the assembled mitochondrial genome. We have used this assembled contig sequence from *Strigamia* as the framework for resequencing the complete mitochondrial genome of this animal. We use this complete sequence and gene order to evaluate whether this mitochondrial genome is useful as a phylogenetic marker for testing ideas about the phylogenetic position of the geophilomorphs within the centipedes, and the centipedes within the wider context of the myriapods and arthropods.

The position of the myriapods within the euarthropods

The relationships between the four euarthropod classes—Chelicerata (arachnids, pycogonids and horse shoe crabs); Crustacea (crabs, copepods etc); Myriapoda (e.g. centipedes and millipedes) and Hexapoda (including insects)—have long been a controversial topic within evolutionary biology. Mitochondrial gene arrangements and molecular phylogenies have convincingly shown that the crustaceans and hexapods form a monophyletic group, the Pancrustacea, in which the hexapods constitute a branch within a larger 'pancrustacean' clade [17]. This well-supported pancrustacean alliance breaks up the old Atelocerata/Uniramia group of Hexapoda and Myriapoda, and the most contentious remaining issue concerns the position of the Myriapoda relative to Pancrustacea and Chelicerata [18]. The traditional grouping of myriapods, hexapods and crustaceans into a group termed the Mandibulata is most obviously based on their shared morphological feature of the post-tritocerebral appendage forming the mandible; chelicerates lack a mandible, and the homologous segment has a pair of walking legs [18]. In contrast, phylogenies compiled from a range of molecular data have tended instead to unite myriapods with the chelicerates in the Myriochelata, rather than to the other mandibulates. [18–20]. Incongruence of morphology and some molecular data have prompted a number of careful studies of the data leading to the suggestion that the support for a Myriochelata grouping may have arisen as a result of systematic error [18]. Resolving this through careful outgroup selection [19] and removing genes with a high rate of nonsynonymous change [21] demonstrated that the strongest phylogenetic signals were in fact in support of Mandibulata. This evidence, and additional analyses of molecular data, indicates a degree of support for Myriapoda as the sister group to Pancrustacea, within a monophyletic Mandibulata [19, 22]. Despite this, the position of the Myriapoda within the Arthropoda remains difficult to resolve.

The position of the chilopods within the myriapods

Extant myriapods are represented by two main groups: the herbivorous millipedes (Diplopoda), and the carnivorous centipedes (Chilopoda). In addition there are two minor groupings: Symphyla and Pauropoda. Whilst the monophyly of each of the four myriapod groups is well-supported by both molecular and morphological studies [23], the inter-relationships of the myriapod classes remain difficult to resolve [24]. Morphological and developmental evidence has traditionally placed the Pauropoda and Diplopoda together as sister lineages in the Dignatha; Symphyla and Dignatha together have been classified as the Progoneata, named for the common presence of an anterior gonopore, with Chilopoda as sister group. In contrast to this, molecular analyses have instead indicated a sister clade relationship between Symphyla and Pauropoda, together forming the Edafopoda [25–27]. Both morphological [28] and molecular [24] studies have yielded a degree of support for a paraphyletic Myriapoda, placing the Chilopoda as sister group to the Chelicerata, and Diplopoda as sister group to Chilopoda + Chelicerata. However, a number of molecular analyses demonstrate strong evidence for the monophyly of the myriapods [25, 26, 29]. More recent phylogenomic analyses support a monophyletic Myriapoda, but place the symphylans as sister group to the three other myriapod classes [30, 31].

The position of the geophilomorphs within the chilopods

The Chilopoda comprises approximately 3000 species within five extant orders: Scutigeromorpha, Lithobiomorpha, Craterostigmomorpha, Scolopendromorpha, and the most diverse order, the Geophilomorpha, to which *Strigamia maritima* belongs [32]. Relationships between chilopod clades seem well resolved from morphological characters and molecular data derived predominantly from single nuclear DNA markers [33]. Molecular data sets support the basal split of the Chilopoda into two evolutionary lineages: the Notostigmophora (= Scutigeromorpha) and Pleurostigmomorpha (the remaining four orders including geophilomorphs), and do not support the alternative hypothesis that Geophilomorpha are the sister group to all other chilopod orders [32–34].

In this study, we describe the complete mitochondrial genome of the centipede *Strigamia maritima*. No geophilomorph mitochondrial genome has been published to date. Here we analyse the gene content and gene order of the *S. maritima* mitochondrial genome in comparison to other arthropod species, and describe the results of a phylogenetic analysis using sequence alignments from mitochondrial protein-coding genes.

Materials and Methods

Initial Sequence from genome scaffold

Within the *S. maritima* whole genome sequence, the scaffold scf718000124766, 23.9kb in length, was found by BLAST to contain a series of mitochondrial protein-coding genes. Closer examination showed atypical large non-coding regions at each end of the scaffold and multiple frameshift errors within protein-coding genes, probably the results of assembly errors within the scaffold. In order to correct possible errors both of assembly and of single mis-read nucleotides we designed PCR primers covering most of the length of the scaffold sequence, and in particular covering all areas containing apparent frameshifts.

DNA Extraction, Primer Design and PCR

DNA was isolated from a population of *Strigamia maritima* living in the wild on the East coast of Scotland [5] and provided to us by the Akam lab. The DNA used came from a pooled sample

of animals, and all sequencing was carried out directly on PCR fragments amplified from this pool. In cases where there is heterozygosity in the population, therefore, the sequence we report will show the most frequently occurring alleles in the PCR product which is likely to represent the highest frequency allele in the population used. Centipedes are not regulated in directive 2010/63/EU of the European Parliament or the UK Animals (Scientific Procedures) Act 1986, but care was taken to minimise potential suffering of the animals.

PCR primers were designed using Primer3 [35] based on the initial 23.9kb scaffold, with the objectives of: i) verifying total genome length, ii) linking the two ends of the mitochondrial genome to produce a closed circle, and iii) correcting sequencing errors within the scaffold sequence. Primer pair sequences are available in the supporting information (S1 Table). Outward facing PCR primers were first designed within conserved gene regions at either end of the scaffold to link both ends of the sequenced genome (to 'close' the circular genome). Within the resulting circular genome (corrected length 14,638 base pairs), PCR primers were designed to amplify the entire sequence in nine overlapping fragments of approximately 2kb each. Where possible, primers were located within conserved protein-coding gene sequences. Of these nine fragments, all but one were successfully amplified. Following gene annotation of the new DNA sequence, likely erroneous stop codons were identified remaining within the coding sequence of *nad6*. New primers were designed to amplify this region allowing us to correct these remaining errors.

All PCRs were performed using the GeneAmp PCR System 2700 (Applied Biosystems, California, USA). PCRs were carried out using the Expand Long-Range PCR Kit (Roche Life Sciences, Penzberg, Germany), following manufacture's recommendations for a 50μl reaction set-up. The Expand Long-Range kit was used owing to its optimisation for amplification of long PCR products, and the high proofreading activity of the polymerase. Cycling was set up as follows: 92°C for 2 min (initial denaturation); 15 cycles of: 92°C for 10 sec (denaturation); 57°C for 15 sec (annealing), 68°C at initial elongation time (approximated as 1 min per 1000 nucleotides to be amplified); 2 cycles each of: 92°C for 10 sec (denaturation), 57°C for 15 sec (annealing), 68°C at 40 sec longer than the initial elongation time, repeated at elongation times increasing by 40 sec intervals for a total of 14 further cycles (two cycles each at seven increasing elongation times). A final elongation stage at 68°C for 7 min was followed by a 4°C 'hold' stage. Amplified products were size separated on ethidium-bromide stained TAE 1% agarose gel and visualised. Successfully amplified products were purified using the High Pure PCR Product Purification Kit (Roche Life Sciences, Penzberg, Germany) with the manufacturer-recommended protocol and sequenced using fluorescent sanger sequencing. Only amplifications which resulted in a single strong band on the agarose gel were purified and sequenced.

Data Assembly and Gene Annotation

For all successful PCR amplifications, forward sequencing results, and the reverse complement of reverse sequencing results, were merged together using the EMBOSS 6.3.1 DNA merger program (http://bioinfo.nhri.org.tw/cgi-bin/emboss/merger), to produce the whole sequenced fragment. The sequence of the amplified closed genome fragment replaced the sequence originally found in the end 4kb and front 3kb regions to correct the length of the mitochondrial genome. The resulting 14,638 base pair circular genome was then used as a point of reference to align each of the sequencing results for the ~2kb fragments (I-VII and IX), resolving the final *Strigamia* genome as 14,983 base pairs in length. Our sequencing results for each of the genes covered by these fragments were compared to the initial sequence to correct any remaining frameshifts, with subsequent results from the *NADH dehydrogenase subunit 6* (*nad6*) fragment fixing the remaining frameshift mutations within this gene. A new consensus sequence for

each gene, based on these results, was generated. In the case of nucleotide ambiguity between the original assembled sequence and new sequencing results, the new sequencing results took preference.

Phylogenetic Inference

Phylogenetic analyses were carried out using a concatenated amino acid alignment of all thirteen protein-coding genes from the *S. maritima* mitochondrial genome. The *S. maritima* protein-coding sequences were first translated using the standard invertebrate mitochondrial genetic code, and the amino acid sequences for each gene were aligned to orthologs from other taxa using MUSCLE [36] (S2 Table). The resulting alignments were trimmed using trimA1 1.2rev59 (with standard settings) [37], and the alignments finally concatenated to produce an alignment of 3407 amino acids from 54 species.

Bayesian analyses was carried out using the site-heterogeneous CAT-GTR mixture model in the PhyloBayes 3.3f software package [38] to allow site-specific amino acid preferences. Four discrete gamma categories are used to distinguish between site-specific rate heterogeneity across the sequence. This model is implemented within a Monte Carlo Markov Chain (MCMC) algorithm, using PhyloBayes. For each alignment, two independent runs were performed for >14,000 cycles and the summary tree calculated with a 'burn in' of 3000 cycles.

Trees were also reconstructed using the maximum likelihood approach using PhyML v 3.0. [39]. The MTArt substitution model was selected, the proportion of invariable sites was estimated and a gamma distribution with 4 categories used. An approximate likelihood ratio test using SH-like supports was conducted to provide estimates of support for clades on the best tree.

Results

Organisation of the genome and genes

The circular, double stranded mitochondrial genome of *S. maritima* is 14,983 base pairs long: 8,925 base pairs shorter than the original contig from the genome assembly (Fig. 1, Table 1). The erroneous additional bases in the original scaffold showed similarity to TY1/Copia-like retrotransposons and our success in closing the circular molecule demonstrates that these derive from incorrect assembly. Our re-sequencing allowed us to correct 15 frame shifts and to correct 584 other incorrectly identified and/or missing nucleotides. The genome contains both the small and large subunit of ribosomal RNA (*rrnS* and *rrnL*), thirteen protein-coding genes (*cytochrome c oxidase* (*cox*) *1*, *2*, *3*; *apocytochrome b* (*cob*); *NADH dehydrogenase* (*nad*) 1, *2*, *3*, *4*, *4l*, 5, *6*; and *ATP synthase F0* (*atp*) 6 and 8) and two large non-coding regions. Sixteen tRNAs were identified using the MiTFi program within MITOS [40, 41]: *trnG*, *trnF*, *trnH*, *trnP*, *trnD*, *trnR*, *trnE*, *trnT*, *trnM*, *trnI*, *trnY*, *trnV*, *trnS2*, *trnN*, *trnK*, *trnL2*. Sequences resembling *trnW*, *trnQ* and *trnA* could only be predicted in the same sequence as *trnI*, *trnE* and *trnT*, respectively, and with low e-values. No credible sequence could be predicted for *trnC*, *trnL1* or *trnS1* using MiTFi or any alternative tRNA prediction software (ARWEN [42], tRNAscan-SE [43]). The predicted sequence for *trnG* is found entirely within the sequence for *cox3*, on the same strand; *trnH* and *trnP* also have partial overlap on the same strand with *nad5* and *nad4l*, respectively. Whilst the predicted sequence of *trnL2* overlaps largely with *rrnL*, they are found on opposite strands (Table 1).

Secondary structures were determined for the sixteen tRNAs which could be reliably identified using MiTFi (Fig. 2). Clear deviations from the classical 'cloverleaf' tRNA secondary structure are observed in many of the *S. maritima* tRNA putative secondary structures. One or more of the four loops are commonly totally or partially missing. The DHU loop is entirely

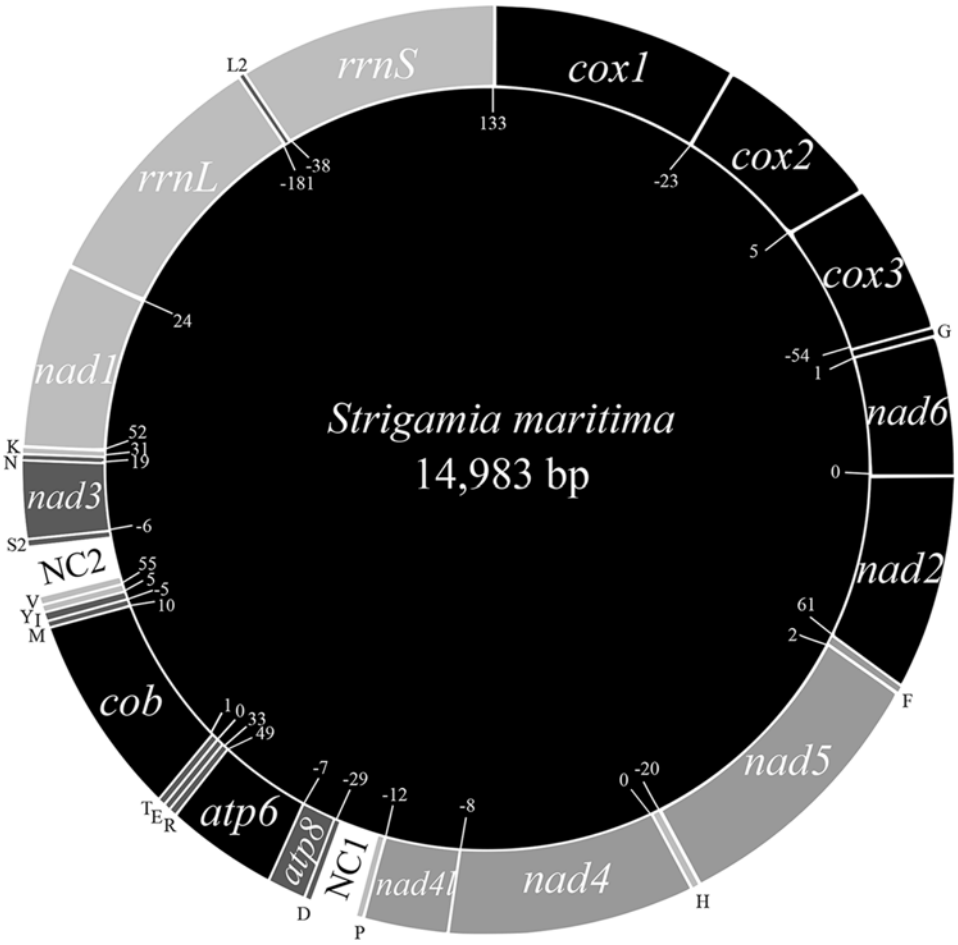


Fig 1. Overview of the mitochondrial genome of *Strigamia maritima* (Myriapoda: Chilopoda), genes not drawn to scale. Numbers inside the circle show intergenic spaces (positive values) or intergenic overlaps (negative values). Protein-coding genes are denoted by three letter abbreviations, ribosomal genes denoted by four letter abbreviations. tRNAs are indicated by single uppercase letters.

doi:10.1371/journal.pone.0121369.g001

missing in *trnR*, and the TΨC loop entirely missing in the predicted structure for *trnN*, *trnE*, *trnG*, *trnM*, *trnS2* and *trnT*. The structure of the acceptor stem is severely truncated or entirely lacking in *trnR*, *trnN*, *trnD*, *trnE*, *trnG*, *trnI*, *trnK* and *trnT*. Within their predicted secondary structures, all tRNAs appear to have mismatched nucleotides and a combination of enlarged or shrunken loops and/or truncated stems.

Mitochondrial genes are transcribed from both strands, with the strand bearing the most protein-coding sequences designated the 'plus' strand. *cox1–3*, *nad6*, *nad2*, *atp8*, *atp6*, *nad3*, as well as *trnD*, *trnR*, *trnE*, *trnT*, *trnM*, *trnI*, *trnS2* and *trnN* are on the plus strand; the five remaining protein-coding genes, both ribosomal genes, and the remaining identified tRNAs, are on the minus strand. Coding DNA accounts for 92.1% of the genome. Of this, and taking into account overlapping regions, protein-coding genes account for 79.31% of the coding DNA, ribosomal genes for 15.24% and tRNA genes for 6.62%. Initiation codons in all *S. maritima* protein-coding genes are ATN: ATT (x5), ATG (x5), ATA (x2) and ATC (x1). Stop codons for all genes were complete: TAA (x8) and TAG (x5) (Tables 1 and 2). Total codon usage across all protein-coding genes is shown in Table 2.

Table 1. Organisation of the *Strigamia maritima* mitochondrial genome.

Gene	Strand	Start position	End position	Length	Start codon	Stop codon	Intergenic nucleotides
cox1	+	1	1557	1557	ATT	TAA	
cox2	+	1532	2215	684	ATG	TAG	-23
cox3	+	2221	3063	843	ATG	TAA	+5
trnG	+	3010	3058	49			-54
nad6	+	3060	3524	464	ATA	TAG	+1
nad2	+	3525	4487	963	ATT	TAA	0
trnF	-	4547	4603	57			+61
nad5	-	4606	6306	1701	ATG	TAG	+2
trnH	-	6287	6347	61			-20
nad4	-	6348	7664	1317	ATG	TAA	0
nad4l	-	7658	7921	264	ATT	TAA	-8
trnP	-	7909	7972	64			-13
NC1		7973	8414	442			0
trnD	+	8415	8489	75			0
atp8	+	8461	8622	162	ATA	TAA	-29
atp6	+	8616	9281	666	ATG	TAA	-7
trnR	+	9331	9375	45			+49
trnE	+	9409	9454	46			+33
trnT	+	9455	9506	52			0
cob	+	9508	10641	1134	ATC	TAG	+1
trnM	+	10652	10710	59			+10
trnI	+	10706	10759	54			-5
trnY	-	10765	10825	61			+5
trnV	-	10880	10937	58			+55
NC2		10938	11331	394			0
trnS2	+	11332	11387	56			0
nad3	+	11382	11732	351	ATT	TAA	-6
trnN	+	11752	11799	48			+19
trnK	-	11831	11888	58			+31
nad1	-	11942	12862	921	ATT	TAG	+52
rrnL	-	12888	14258	1371			+24
trnL2	+	14078	14148	71			-181
rrnS	-	14111	14850	740			-38

Intergenic nucleotides shown as gaps (positive values) or overlap (negative values) between consecutive genes.

doi:10.1371/journal.pone.0121369.t001

The A+T content of the total genome is 64.02%, which is lower than the percentage found in the mitochondrial genome of the centipedes *Lithobius forficatus* (67.9%) [44] and *Scutigera coleoptrata* (69.4%)[45], the pauropod *Pauropus longiramus* (72.9%) [27], the symphylan *Scutigerella causeyae* (72.6%)[20] and the millipede *Thyropygus* sp. (67.8%) [46]; close to that of the millipede *Narceus annularis* (63.7%) [46]; and higher than that of the millipede *Antrokoreana gracilipes* (62.1%) [47]. The A+T content of the coding portion of the genome is 63.5%. Average A+T content across the thirteen protein-coding genes is 62.8%; lower than the 68.58% average A+T content of the two ribosomal genes. As a result of the high A+T content of the mitochondrial genome, the most frequently occurring codons across the protein-coding genes are those comprised of A and T nucleotides: TTT (x 226), ATT (x 195), TTA (x 184)

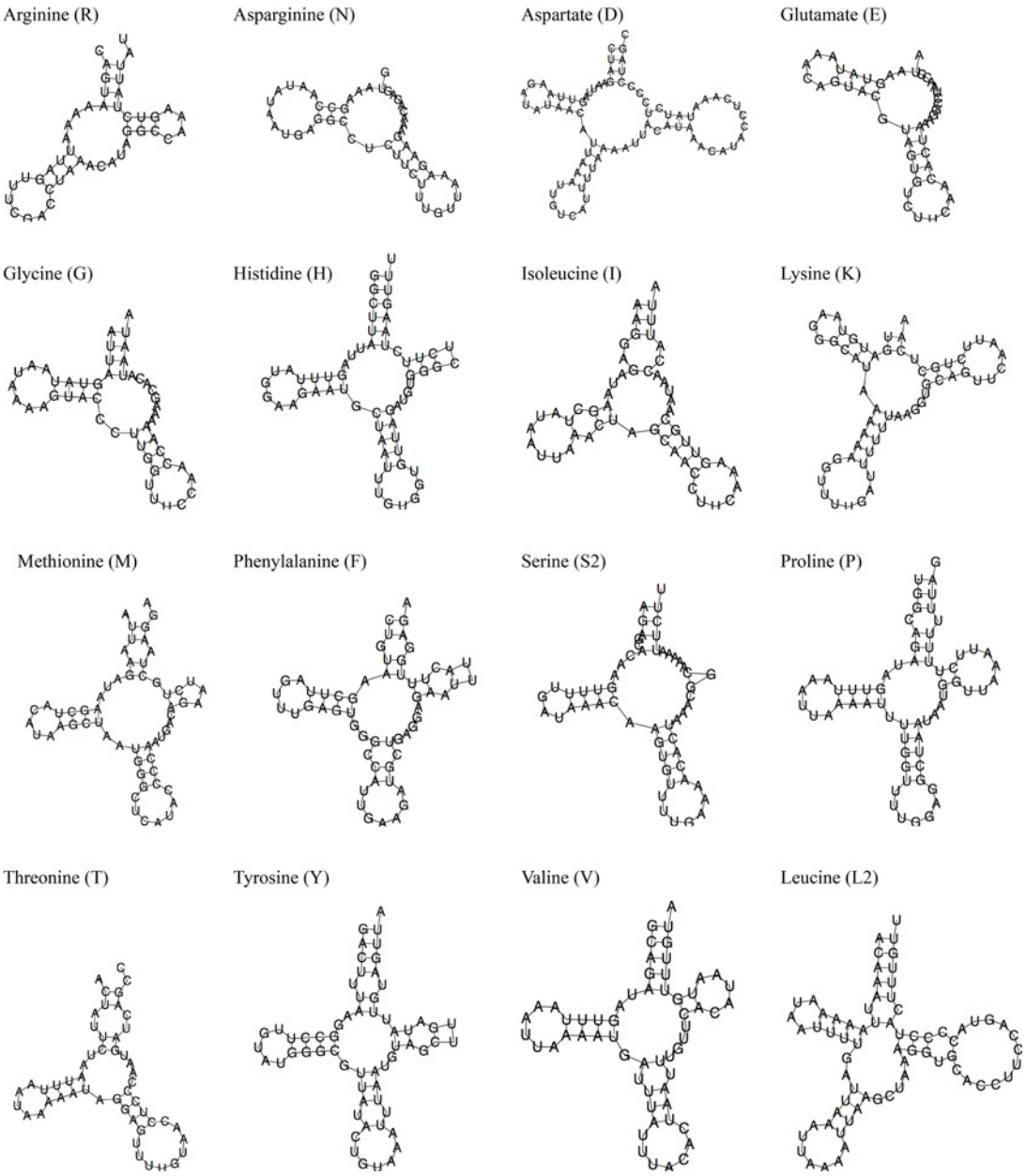


Fig 2. Putative secondary structures of tRNAs from the mitochondrial genome of *Strigamia maritima* as predicted by MiTFi [40, 41].

doi:10.1371/journal.pone.0121369.g002

and ATA (x 169) (Table 2). Two main non-coding regions, NC1 and NC2 (Table 1) were found to have a higher A+T content than that of the total genome: 71.27% for NC1 and 71.07% for NC2. The compositional difference between the non-coding regions and the genome as a whole is statistically significant (NC1, $\chi 2 = 9.503$, $p<0.01$; NC2, $\chi 2 = 7.991$, $p<0.01$); consequently, these two regions are proposed as control regions [48]. Eight other short non-coding regions, ranging in length from 19 to 131 nucleotides are also found throughout the genome. None of these regions have an A+T content that is statistically significantly higher than that of the genome as a whole.

Table 2. Total codon usage across the protein-coding genes of the *Strigamia* mitochondrial genome.

Codon	Aa	Frequency	Codon	Aa	Frequency	Codon	Aa	Frequency	Codon	Aa	Frequency
TTT	F	226	**TCT**	S	96	**TAT**	Y	82	**TGT**	C	25
TTC	F	61	**TCC**	S	31	**TAC**	Y	53	**TGC**	C	16
TTA	L	184	**TCA**	S	71	**TAA**	STOP	8	**TGA**	W	71
TTG	L	82	**TCG**	S	10	**TAG**	STOP	5	**TGG**	W	35
CTT	L	81	**CCT**	P	44	**CAT**	H	39	**CGT**	R	12
CTC	L	23	**CCC**	P	54	**CAC**	H	43	**CGC**	R	5
CTA	L	136	**CCA**	P	59	**CAA**	Q	61	**CGA**	R	35
CTG	L	19	**CCG**	P	6	**CAG**	Q	15	**CGG**	R	10
ATT	I	195	**ACT**	T	41	**AAT**	N	47	**AGT**	S	25
ATC	I	91	**ACC**	T	55	**AAC**	N	56	**AGC**	S	15
ATA	M	169	**ACA**	T	129	**AAA**	K	65	**AGA**	S	60
ATG	M	57	**ACG**	T	8	**AAG**	K	17	**AGG**	S	15
GTT	V	117	**GCT**	A	62	**GAT**	D	32	**GGT**	G	59
GTC	V	16	**GCC**	A	57	**GAC**	D	35	**GGC**	G	47
GTA	V	79	**GCA**	A	83	**GAA**	E	48	**GGA**	G	67
GTG	V	53	**GCG**	A	14	**GAG**	E	37	**GGG**	G	122

Standard one-letter abbreviations for amino acids (Aa) using invertebrate mitochondrial genetic code.

doi:10.1371/journal.pone.0121369.t002

Nucleotide composition of the plus strand is as follows: A = 38.96%; T = 25.06%, C = 23.90% and G = 12.08%. Base compositional bias between the two strands can be measured as GC- and AT- skew, where GC-skew = (G—C)/(G + C) and AT-skew = (A—T)/(A + T)[49]. Using these formulae, skew values are generated ranging in value from-1 to +1; an absolute value closer to 1 indicates compositional asymmetry between the two stands, whilst a value of 0 indicates that distribution is equal between the strands. For the *S. maritima* plus strand, GC-skew = -0.33 and AT-skew = 0.22, showing asymmetry in nucleotide composition between the two strands. The absolute GC-skew value is higher than that found in *Thyropygus sp* (-0.29) [46], *L. forficatus* (-0.27) [44] and *S. coleoptrata* (-0.31) [45], but lower than that of *N. annularis* (-0.40) [46]. Absolute AT-skew is higher than that in *Thyropygus sp* (0.08) [46], *L. forficatus* (0.09) [44], *S. coleoptrata* (0.04) [45] *N. annularis* (0.07) [46] and *S. causeyae* (-0.12) [20].

Gene order

The overall arrangement of genes around the *S. maritima* mitochondrial genome is unique compared to other arthropod species or to any other metazoan mitochondrial genome studied (Fig. 3). Genes of the same transcriptional polarity are clustered together, with the exception of *trnL2*, which overlaps with genes on the opposite strand. Four blocks of protein-coding genes follow the arthropod 'ground plan' (Fig. 3): *cox1-cox2* (plus strand); *trnF-nad5-trnH-nad4-nad4l-trnP* (minus strand); *trnD-atp8-atp6* (translocated towards the 3' end of the plus strand); and *nad1-rrnL-rrnS* (translocated towards the 5' end of the minus strand). The composition of the rest of the genome has a gene order which is completely unique to *Strigamia*: *nad6* and *nad2* have been rearranged adjacent to the *cox1-cox2-trnG-cox3* block at the 5' end of the plus strand; *cob* has rearranged to the 3' end of the *trnD-atp8-atp6* block, with the addition of *trnR*, *trnE*, *trnT* on the 5' side, and *trnM* and *trnI* on the 3' end; and *trnS2-nad3-trnN* is a novel arrangement on the minus strand. Two main non-coding control regions are proposed, one between *trnP* and *trnD*, and the other between *trnV* and *trnS*2. The location of these is unique to

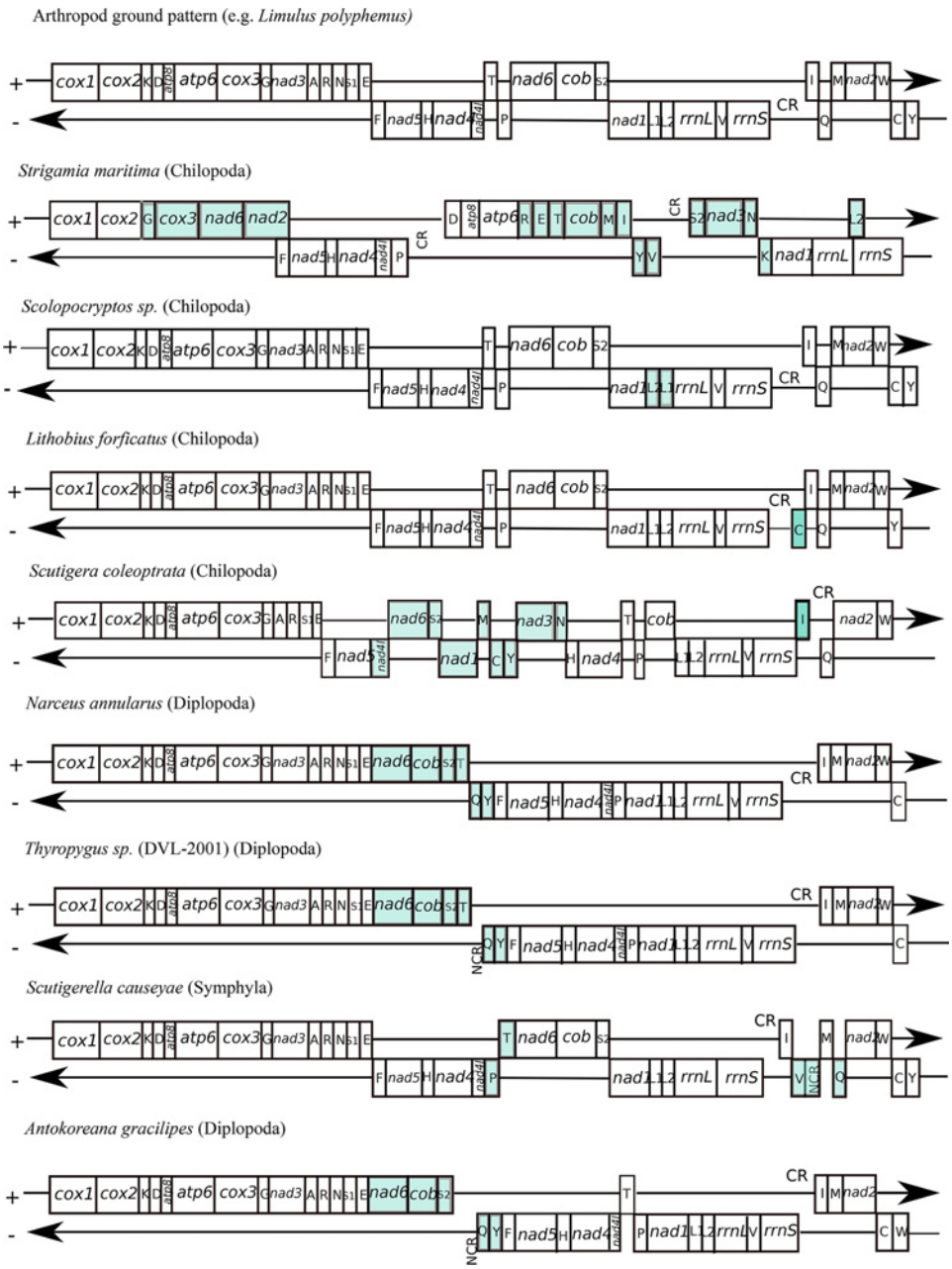


Fig 3. Comparisons of gene orders in the mitochondrial genomes of *Strigamia maritima*, the arthropod 'ground-plan' (*Limulus polyphemus*), three further chilopods, three diplopods and a symphylan. Genes are not drawn to scale; shaded genes indicate those that have moved from the original arthropod 'ground-plan'.

doi:10.1371/journal.pone.0121369.g003

Strigamia. Compared to other arthropod species, the genes of *S. maritima* have a large degree of overlap and of intergenic space (Table 1).

Phylogenetic Analysis

Using a data set of 54 species, PhyloBayes Bayesian and Maximum Likelihood (ML) phylogenetic analysis was performed using conserved blocks of amino acid alignments of protein-coding

genes (Fig. 4). The arthropod portion of the tree is rooted with the deuterostome cephalochordate *Epigonichthys lucayanus* as well as five lophotrochozoans (two molluscs, one annelid and two brachiopods) and the ecdysozoan priapulid *Priapulus caudatus*. In the Bayesian phylogeny, (Fig. 4) Myriapoda and Chelicerata, 'Myriochelata', are resolved as the sister group to Pancrustacea, with Bayesian Posterior Probabilities (BPP) of 0.94 (Myriochelata) and BPP = 1 (Pancrustacea). Within the clade of Myriochelata the chelicerates are supported as monophyletic with maximum support, but the analysis did not resolve the myriapods as monophyletic. Three myriapod clades are resolved, however: Chilopoda (BPP = 0.99); Diplopoda plus Pauropoda (BPP = 0.53; Diplopoda alone BPP = 0.98) and Symphyla (BPP = 1). Our ML analysis (Fig. 5) resolves a monophyletic Mandibulata (apart from the anomalous pauropod) but shows a paraphyletic Myriapoda, placing the Chilopoda as sister group to Crustacea + Hexapoda with SH-like support value 0.99, and the Pauropoda (represented by *Pauropus longiramus*) as sister group to the Chelicerata (SH-like support = 0.97). The internal relationships of the four Chilopoda orders in both of our phylogenetic analyses corroborates the consensus opinion on centipede relationships derived from other molecular analyses [33].

Discussion

Genome composition and tRNAs

All thirteen protein-coding genes and both ribosomal genes were found in the *S. maritima* mitochondrial genome. Only sixteen of the standard 22 tRNAs were identified. The lack of detectable tRNAs using our bionformatic analysis may be due to a truncation of tRNA sequences and/or asymmetry of tRNA secondary structure such as have been previously found in other

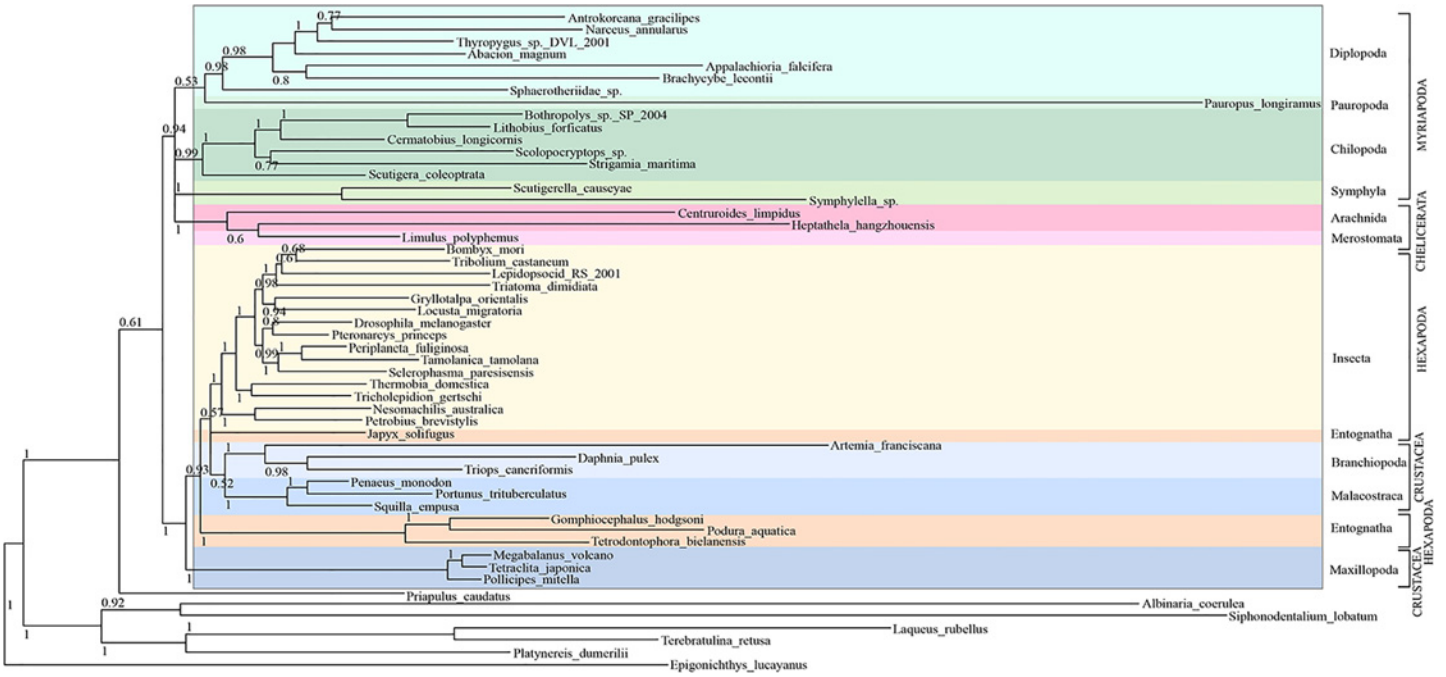


Fig 4. Bayesian phylogenetic analysis of mitochondrial protein-coding genes from Arthropoda species including *S. maritima*. Support values at nodes are Bayesian Posterior Probability (BPP). Myriapoda and Chelicerata, 'Myriochelata' (BPP = 0.94) resolved as the sister group to Pancrustacea (Crustacea and Hexapoda, BPP = 1.0). A monophyletic Chilopoda is resolved with BPP = 0.99, within which Scutigeromorpha (*Scutigera coleoptrata*) are resolved as the sister group to the three remaining chilopod orders represented in our phylogeny (Lithobiomorpha, (*Bothropolys sp.*, *Lithobius forficatus* and *Cermatobius longicornis*); Scolopendromorpha (*Scolopocryptos sp.*) and Geophilomorpha (*Strigamia maritima*)) with BPP = 1.

doi:10.1371/journal.pone.0121369.g004

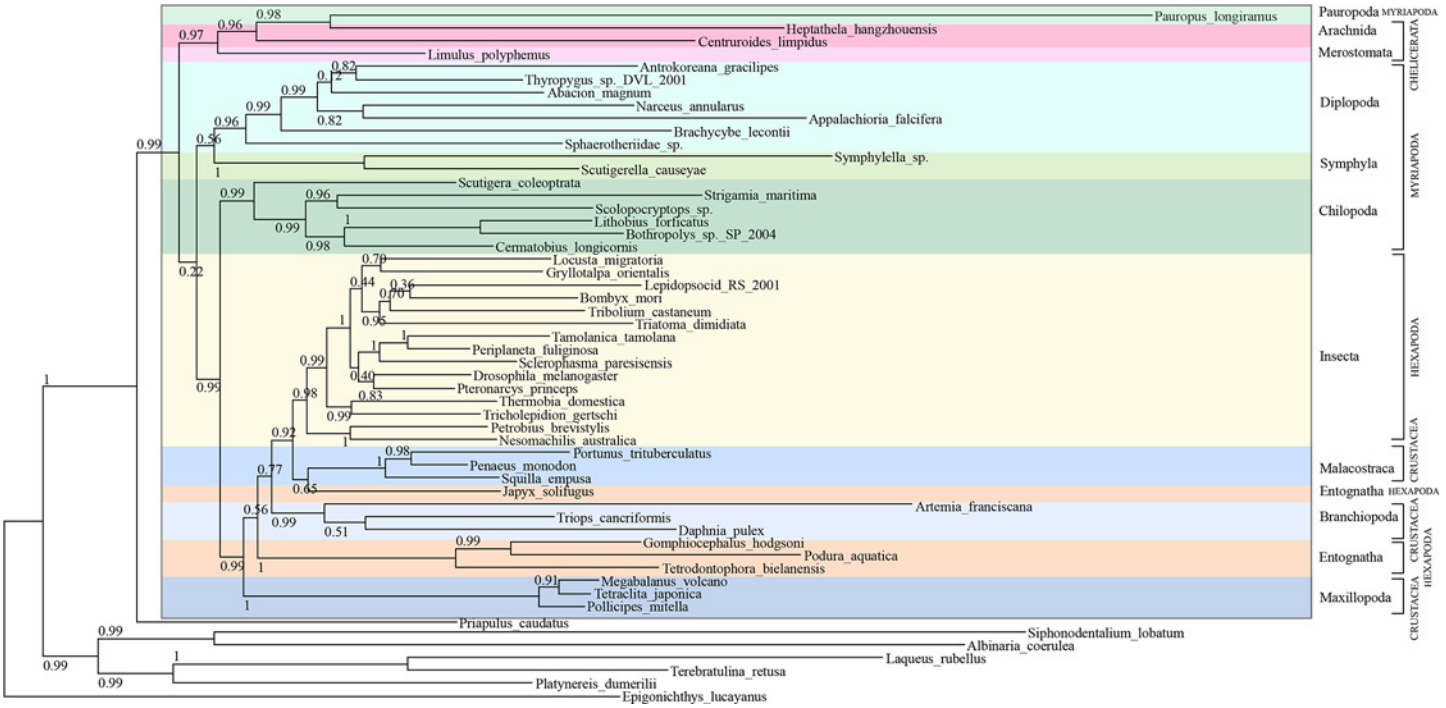


Fig 5. Maximum Likelihood phylogenetic analysis of mitochondrial protein-coding genes from Arthropoda species including *S. maritima*. Support at nodes are SH-like support values. A monophyletic Chilopoda is resolved as sister group to Pancrustacea (SH-like support = 0.99) and Pauropoda (represented by *Pauropus longiramus*) placed as sister group to Chelicerata (SH-like support = 0.97). Within the Chilopoda, Scutigeromorpha (*Scutigera coleoptrata*) are resolved as sister group to the three other chilopod orders represented in our phylogeny (Lithobiomorpha, (*Bothropolys sp*., *Lithobius forficatus* and *Cermatobius longicornis*); Scolopendromorpha (*Scolopocryptos sp*.) and Geophilomorpha (*Strigamia maritima*) with SH-like support = 0.99.

doi:10.1371/journal.pone.0121369.g005

arthropod genomes [20, 27]. *trnW*, *trnQ* and *trnA* could only be predicted within the same sequence as *trnI*, *trnE* and *trnT*, respectively, which demonstrates a degree of ambiguity in determining tRNA sequences for *Strigamia* using bioinformatic approaches.

For the sixteen tRNA sequences that could be identified, it is clear that they do not all conform to the canonical cloverleaf-shaped secondary structure of tRNAs: for all tRNAs, stems contain mismatches and/or are truncated, and one or more loop may be either missing or greatly modified (Fig. 3). This situation is not unique to *Strigamia*: complete loss of the TΨC loop in tRNA has been described in a number of metazoan mitochondrial genomes, including the jumping spider *Habronattus oregonensis* (Chelicerata) [50]. As in *Strigamia*, *H. oregonensis* has asymmetric tRNA sequences which cannot be folded into a typical cloverleaf-secondary structure [50]. In addition, many of its tRNAs also lack a fully paired acceptor stem. In both *Strigamia* and *Habronattus*, tRNAs overlap with other RNAs or protein-coding genes on the same or the opposite strand, which could result in a truncation of what would normally be the acceptor stem of the tRNA. It is also possible that the 3' portion of the acceptor stem may be formed post-transcriptionally, as in the centipede *Lithobius forficatus* [44]. Truncated tRNAs as a result of overlapping genes may be the result of a tendency to reduce mitochondrial genome size, as has also been proposed for the myriapod *Pauropus longiramus* [27], but the evolutionary advantage of this is uncertain. One proposed outcome of incomplete tRNAs is that they cause an accumulation of deleterious mutations at a faster-than-normal rate, leading to a potential 'mutational meltdown'[51]. Theoretically, if posttranscriptional modification could keep up with the accumulation of mutations, as well as reduce the mitochondrial genome size, the truncated tRNAs would be retained whilst reducing the mitochondrial genome size as observed.

Gene Order

Previous comparisons across the arthropods, and more widely within the Ecdysozoa, proposed that the ancestral arthropod mitochondrial genome has a gene arrangement identical to that found in *Limulus polyphemus* [52]. Mitochondrial gene order can be an informative phylogenetic tool, and a significant finding of this study is that gene order in *S. maritima* is notably different from that of any other myriapod, or indeed any other metazoan species, to which it can be compared. Whilst small regions of gene order in *Strigamia* follow that of the arthropod 'ground pattern', (for example, *trnF-nad5-trnH-nad4-nad4L* on the minus strand), other sections are completely rearranged without a precedent among metazoans.

Gene order rearrangement is most commonly thought to occur via a 'duplication and deletion' model. This proposes that the random duplication of part of the mitochondrial genome occurs as a result of slipped-strand mispairing or an error during replication termination. Following this, one of the gene copies is deleted. If it is the original copy that becomes deleted this results in a change in gene order [53]. Evidence for this model is provided by mitochondrial genomes with duplicated regions including at least one protein-coding or rRNA gene [53, 54].

In a recent study concerning the house centipede *Scutigera coleoptrata* (Scutigeromorpha), a novel mitochondrial gene arrangement could only be explained by postulating as many as 10 gene translocations and/or duplications and losses involving four protein-encoding genes (*nad3*,*nad4L*, *nad6*, and *nad1*) and six tRNAs genes (*trnN*, *trnS2*, *trnL*, *trnM*, *trnC* and *trnY*) [45]. Gene rearrangement in *S. maritima* is not as easily accommodated by the duplication, loss and translocation theory of gene rearrangement. To derive the *Strigamia maritima* mitochondrial gene order from the *Limulus polyphemus* 'ground plan' would require gene translocations involving five protein-coding genes (*cox3*, *cob*, *nad6*, *nad2* and *nad3*) and eleven tRNAs (*trnR*, *trnE*, *trnT*, *trnM*, *trnI*, *trnL2*, *trnY*, *trnV*, *trnS*, *trnN*, and *trnK*), and the observed order of these is not easily reached from the ancestral arrangement.

Alternative mechanisms for gene rearrangement may therefore be necessary in order to explain the gene order observed in *Strigamia*. As outlined, identifying tRNA sequences using computational analysis was made difficult by the asymmetry of their secondary stem and loop structure. A possible explanation for the novel gene order is that the mechanism for gene rearrangement in *Strigamia* relies on stem and loop structures [55, 56]. In vertebrates, the endpoints of tandemly duplicated gene regions contain stem and loop structures: either from tRNAs or from the protein-coding gene regions [57]. More widely, in both vertebrates and invertebrates, tRNA genes are involved more frequently in mitochondrial gene rearrangements than protein-coding or ribosomal genes [27, 55]. It is possible that the asymmetrical and truncated structure of the *Strigamia* tRNAs leads to randomly located tandem repeats occurring simultaneously at many locations along the mitochondrial genome. Extensive gene rearrangement could alternatively be explained by small direct repeats. In ranid frogs, transpositions in terminal inverted or direct repeats have created non-functional copies of *trnL2* in the same position as the functional copy. Transposition of the repeated copy has subsequently resulted in a copy of trnL2 that is 5kb away from the 'usual' position observed in other vertebrates [56]. This pattern of rearrangement is similar to that observed in *Strigamia*, where *trnL2* is inverted from the ancestral position and has been translocated into the coding region, overlapping with *rrnS* and *rrnL*.

In our phylogenetic analysis, a sister-group relationship is weakly supported between Geophilomorpha (*Strigamia maritima*) and Scolopendromorpha (*Scolopocryptos sp.*) The mitochondrial gene order of *Scolopocryptos sp* has recently been shown as identical to that of the arthropod 'ground plan' represented by *Limulus polyphemus*, except for the interchanged positions of *trnL1* and *trnL2* [58] (Fig. 3). It appears, therefore, that no meaningful phylogenetic

information can be derived from comparing the gene order of these two species, as any differences would be *Strigamia* specific. Further sequencing of mitochondrial genomes from additional members of the Geophilomorpha would show whether such extensive rearrangement is unique to *Strigamia* and hence a recent innovation or found commonly throughout this order of centipedes and hence a more ancient event. It is also apparent that the novel gene order found in the *Strigamia* mitochondrial genome contrasts with the exceptionally conservative gene content and arrangement observed in its nuclear genome [16].

Phylogenetic Inference

Resolving the inter-relatedness of the Myriapoda, and determining their position within the Arthropoda, remains a difficult phylogenetic problem. In our Bayesian phylogeny (Fig. 4), the myriapods form a poorly resolved clade with the chelicerates, thus favouring the Myriochelata and Pancrustacea hypothesis over that of a monophyletic Mandibulata [19]. Our ML analysis (Fig. 5) resolves a paraphyletic Myriapoda, placing Chilopoda as sister group to Crustacea + Hexapoda, separate from the Diplopoda + Symphyla grouping. Pauropoda are resolved as sister group to the Chelicerata. SH-like branch support for the node splitting off (Diplopoda + Symphyla (Chilopoda (Crustacea + Hexapoda))) is only very low [0.22]. Phylogenies derived from mitochondrial DNA of other arthropod members have also resolved Myriochelata [20, 59, 60], but this is not always a well-supported grouping [20]. As mitochondrial DNA has a high A+T content—averaging approximately 70% in metazoan taxa—the likelihood of compositional bias and multiple substitutions means that phylogenies derived from mitochondrial genes are particularly prone to systematic error [24, 61]. Ecdysozoan phylogenies based on a much larger set of nuclear genes support a monophyletic Mandibulata and monophyletic Myriapoda once systematic error has been carefully dealt with [62]. It therefore seems probable that the data we have in this analysis is too small a sample, as well as possibly suffering from systematic error due to obvious compositional biases, to reconstruct these relationships accurately.

In this study, the centipedes, Chilopoda, are resolved as a monophyletic grouping, and the relationships within the order correspond with those derived from previous molecular analyses of nuclear ribosomal and nuclear protein-coding genes [33]. Scutigeromorpha, represented by *Scutigera coleoptrata* in our analysis, is found as the sister group to the three remaining centipede orders represented in our phylogeny (Lithobiomorpha, (*Bothropolys sp*., *Lithobius forficatus* and *Cermatobius longicornis*); Scolopendromorpha (*Scolopocryptos sp*.) and Geophilomorpha (*Strigamia maritima*)) with BPP = 1 and SH-like support = 0.99. Together with the Craterostigmomorpha—an order from which no mitochondrial genes have been sequenced—these three orders form the Pleurostigmomorpha. Our phylogenies therefore conform to the widely-held view that Scutigeromorpha are the 'sister-order' to the four remaining orders forming the Pleurostigmomorpha [33]. Our phylogenies also support the sister-group relationship between Geophilomorpha (*Strigamia maritima*) and Scolopendromorpha (*Scolopocryptos sp*.) with 0.77 BPP and 0.76 SH-like support in our ML phylogeny.

We sequenced the first complete mitochondrial genome of a geophilomorph centipede. Phylogenetic analyses using mitochondrial protein-coding genes were unable to support a monophyletic Mandibulata, but did support a monophyletic Chilopoda with inter-relatedness conforming to the view that Scutigeromorpha are the sister group to the four remaining chilopod orders comprising the Pleurostigmomorpha. Gene order of the *Strigamia* mitochondrial genome is unique compared to any other arthropod, or indeed any other metazoan, mitochondrial genome studied. This unusual organisation contrasts with the notably conservative nuclear genome [16]. Further sequencing and analysis of mitochondrial genomes from this order of

centipedes is therefore required to see whether this unusual gene order is unique to *Strigamia*, or common to members of the Geophilomorpha.

Supporting Information

S1 Table. Primer pairs used for amplification of fragments within the mitochondrial genome of *Strigamia maritima*.
(DOCX)

S2 Table. GenBank accession numbers for taxa used in this study.
(DOCX)

Acknowledgments

We would like to thank Michael Akam (Department of Zoology, University of Cambridge) and Stephen Richards (Department of Molecular and Human Genetics, Baylor College of Medicine) for their work on the *Strigamia* nuclear genome, and Bernhard Egger and members of the Telford lab for their help in analysis.

Author Contributions

Conceived and designed the experiments: HER FL. Performed the experiments: HER FL. Analyzed the data: HER FL MJT ACR. Wrote the paper: HER FL MJT.

References

1. Barber AD. Littoral myriapods: a review. Soil Organisms. 2009; 81(3):26.
2. Arthur W, Chipman AD. The centipede Strigamia maritima: what it can tell us about the development and evolution of segmentation. Bioessays. 2005; 27(6):653–60. PMID: 15892117
3. Brena C, Akam M. The embryonic development of the centipede Strigamia maritima. Dev Biol. 2012; 363(1):290–307. doi: 10.1016/j.ydbio.2011.11.006 PMID: 22138381
4. JGE L. The life history and ecology of the littoral centipede *Strigamia maritima* (Leach). Proceedings of the Zoological Society of London. 1961; 137:221–48.
5. Chipman AD, Arthur W, Akam M. Early development and segment formation in the centipede, Strigamia maritima (Geophilomorpha). Evol Dev. 2004; 6(2):78–89. PMID: 15009120
6. Chipman AD, Stollewerk A. Specification of neural precursor identity in the geophilomorph centipede Strigamia maritima. Dev Biol. 2006; 290(2):337–50. PMID: 16380110
7. Green JE, Akam M. Germ cells of the centipede Strigamia maritima are specified early in embryonic development. Dev Biol. 2014; 392(2):419–30. doi: 10.1016/j.ydbio.2014.06.003 PMID: 24930702
8. Hunnekuhl VS, Akam M. An anterior medial cell population with an apical-organ-like transcriptional profile that pioneers the central nervous system in the centipede Strigamia maritima. Dev Biol. 2014.
9. Chipman AD, Arthur W, Akam M. A double segment periodicity underlies segment generation in centipede development. Curr Biol. 2004; 14(14):1250–5. PMID: 15268854
10. Brena C, Akam M. An analysis of segmentation dynamics throughout embryogenesis in the centipede Strigamia maritima. BMC Biol. 2013; 11:112. doi: 10.1186/1741-7007-11-112 PMID: 24289308
11. Chipman AD, Akam M. The segmentation cascade in the centipede Strigamia maritima: involvement of the Notch pathway and pair-rule gene homologues. Dev Biol. 2008; 319(1):160–9. doi: 10.1016/j.ydbio.2008.02.038 PMID: 18455712
12. Green J, Akam M. Evolution of the pair rule gene network: Insights from a centipede. Dev Biol. 2013; 382(1):235–45. doi: 10.1016/j.ydbio.2013.06.017 PMID: 23810931
13. Kettle C, Johnstone J, Jowett T, Arthur H, Arthur W. The pattern of segment formation, as revealed by engrailed expression, in a centipede with a variable number of segments. Evol Dev. 2003; 5(2):198–207. PMID: 12622737
14. Brena C, Green J, Akam M. Early embryonic determination of the sexual dimorphism in segment number in geophilomorph centipedes. Evodevo. 2013; 4(1):22. doi: 10.1186/2041-9139-4-22 PMID: 23919293

15. Vedel V, Apostolou Z, Arthur W, Akam M, Brena C. An early temperature-sensitive period for the plasticity of segment number in the centipede Strigamia maritima. Evol Dev. 2010; 12(4):347–52. doi: 10.1111/j.1525-142X.2010.00421.x PMID: 20618430
16. Chipman AD, Ferrier DE, Brena C, Qu J, Hughes DS, Schröder R, et al. The First Myriapod Genome Sequence Reveals Conservative Arthropod Gene Content and Genome Organisation in the Centipede Strigamia maritima. PLoS Biol. 2014; 12(11):e1002005. doi: 10.1371/journal.pbio.1002005 PMID: 25423365
17. Boore JL, Lavrov DV, Brown WM. Gene translocation links insects and crustaceans. Nature. 1998; 392 (6677):667–8. PMID: 9565028
18. Edgecombe GD. Arthropod phylogeny: an overview from the perspectives of morphology, molecular data and the fossil record. Arthropod Struct Dev. 2010; 39(2–3):74–87. doi: 10.1016/j.asd.2010.05.004 PMID: 20566316
19. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, et al. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Biol Sci. 2011; 278(1703):298–306. doi: 10.1098/rspb.2010.0590 PMID: 20702459
20. Podsiadlowski L, Kohlhagen H, Koch M. The complete mitochondrial genome of Scutigerella causeyae (Myriapoda: Symphyla) and the phylogenetic position of Symphyla. Mol Phylogenet Evol. 2007; 45 (1):251–60. PMID: 17764978
21. Regier JC, Shultz JW, Ganley AR, Hussey A, Shi D, Ball B, et al. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst Biol. 2008; 57 (6):920–38. doi: 10.1080/10635150802570791 PMID: 19085333
22. Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature. 1995; 376(6536):163–5. PMID: 7603565
23. Shear WA, Edgecombe GD. The geological record and phylogeny of the Myriapoda. Arthropod Struct Dev. 2010; 39(2–3):174–90. doi: 10.1016/j.asd.2010.05.004 PMID: 20566316
24. Negrisolo E, Minelli A, Valle G. The mitochondrial genome of the house centipede scutigera and the monophyly versus paraphyly of myriapods. Mol Biol Evol. 2004; 21(4):770–80. PMID: 14963096
25. Gai YH, Song DX, Sun HY, Zhou KY. Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences. Zoolog Sci. 2006; 23(12):1101–8. PMID: 17261924
26. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, et al. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature. 2010; 463(7284):1079–83. doi: 10.1038/nature08742 PMID: 20147900
27. Dong Y, Sun H, Guo H, Pan D, Qian C, Hao S, et al. The complete mitochondrial genome of Pauropus longiramus (Myriapoda: Pauropoda): implications on early diversification of the myriapods revealed from comparative analysis. Gene. 2012; 505(1):57–65. doi: 10.1016/j.gene.2012.05.049 PMID: 22659693
28. Loesel R, Nässel DR, Strausfeld NJ. Common design in a unique midline neuropil in the brains of arthropods. Arthropod Structure & Development. 2002; 31(1):77–91.
29. Regier JC, Wilson HM, Shultz JW. Phylogenetic analysis of Myriapoda using three nuclear protein-coding genes. Mol Phylogenet Evol. 2005; 34(1):147–58. PMID: 15579388
30. Miyazawa H, Ueda C, Yahata K, Su ZH. Molecular phylogeny of Myriapoda provides insights into evolutionary patterns of the mode in post-embryonic development. Sci Rep. 2014; 4:4127. doi: 10.1038/srep04127 PMID: 24535281
31. Rehm P, Meusemann K, Borner J, Misof B, Burmester T. Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing. Mol Phylogenet Evol. 2014; 77:25–33. doi: 10.1016/j.ympev.2014.04.007 PMID: 24732681
32. Edgecombe GD, Giribet G. Evolutionary biology of centipedes (Myriapoda: Chilopoda). Annu Rev Entomol. 2007; 52:151–70. PMID: 16872257
33. Edgecombe GD. Centipede systematics: progress and problems. Zootaxa. 2007; 1668:327–41.
34. Edgecombe GD, Giribet G. Adding mitochondrial sequence data (16S rRNA and cytochrome c oxidase subunit I)to the phylogeny of centipedes (Myriapoda:Chilopoda):an analysis of morphologyand four molecular loci. Journal of Zoological Systematics and Evolutionary Research. 2004; 42:89–134.
35. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012; 40(15):e115. PMID: 22730293
36. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5:113. PMID: 15318951

37. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972–3. doi: 10.1093/bioinformatics/btp348 PMID: 19505945

38. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 2009; 25(17):2286–8. doi: 10.1093/bioinformatics/btp368 PMID: 19535536

39. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010; 59(3):307–21. doi: 10.1093/sysbio/syq010 PMID: 20525638

40. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013; 69(2):313–9. doi: 10.1016/j.ympev.2012.08.023 PMID: 22982435

41. Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, et al. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res. 2012; 40(7):2833–45. doi: 10.1093/nar/gkr1131 PMID: 22139921

42. Laslett D, Canbäck B. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics. 2008; 24(2):172–5. PMID: 18033792

43. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33(Web Server issue):W686–9. PMID: 15980563

44. Lavrov DV, Brown WM, Boore JL. A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede Lithobius forficatus. Proc Natl Acad Sci U S A. 2000; 97(25):13738–42. PMID: 11095730

45. Negrisolo E, Minelli A, Valle G. Extensive gene order rearrangement in the mitochondrial genome of the centipede Scutigera coleoptrata. J Mol Evol. 2004; 58(4):413–23. PMID: 15114420

46. Lavrov DV, Boore JL, Brown WM. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. Mol Biol Evol. 2002; 19(2):163–9. PMID: 11801744

47. Woo HJ, Lee YS, Park SJ, Lim JT, Jang KH, Choi EH, et al. Complete mitochondrial genome of a troglobite millipede Antrokoreana gracilipes (Diplopoda, Juliformia, Julida), and juliformian phylogeny. Mol Cells. 2007; 23(2):182–91. PMID: 17464195

48. Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene. 1999; 238(1):195–209. PMID: 10570997

49. Perna NT, Kocher TD. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol. 1995; 41(3):353–8. PMID: 7563121

50. Masta SE, Boore JL. The complete mitochondrial genome sequence of the spider Habronattus oregonensis reveals rearranged and extremely truncated tRNAs. Mol Biol Evol. 2004; 21(5):893–902. PMID: 15014167

51. Gabriel W, Lynch M, Burger R. Muller's Ratchet and mutational meltdowns. Evolution. 1993; 47(6):1744–57.

52. Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999; 27(8):1767–80. PMID: 10101183

53. Boore JL, Brown WM. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev. 1998; 8(6):668–74. PMID: 9914213

54. Moritz C, Brown WM. Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. Proc Natl Acad Sci U S A. 1987; 84(20):7183–7. PMID: 3478691

55. Macey JR, Larson A, Ananjeva NB, Papenfuss TJ. Replication slippage may cause parallel evolution in the secondary structures of mitochondrial transfer RNAs. Mol Biol Evol. 1997; 14(1):30–9. PMID: 9000751

56. Macey JR, Larson A, Ananjeva NB, Fang Z, Papenfuss TJ. Two novel gene orders and the role of light-strand replication in rearrangement of the vertebrate mitochondrial genome. Mol Biol Evol. 1997; 14(1):91–104. PMID: 9000757

57. Stanton DJ, Daehler LL, Moritz CC, Brown WM. Sequences with the potential to form stem-and-loop structures are associated with coding-region duplications in animal mitochondrial DNA. Genetics. 1994; 137(1):233–41. PMID: 8056313

58. Gai Y, Ma H, Ma J, Li C, Yang Q. The complete mitochondrial genome of Scolopocryptops sp. (Chilopoda: Scolopendromorpha: Scolopocryptopidae). Mitochondrial DNA. 2014; 25(3):192–3. doi: 10.3109/19401736.2013.792073 PMID: 23631366

59. Pisani D, Poling LL, Lyons-Weiler M, Hedges SB. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol. 2004; 2:1. PMID: 14731304

60. Gai Y, Song D, Sun H, Yang Q, Zhou K. The complete mitochondrial genome of Symphylella sp. (Myriapoda: Symphyla): Extensive gene order rearrangement and evidence in favor of Progoneata. Mol Phylogenet Evol. 2008; 49(2):574–85. doi: 10.1016/j.ympev.2008.08.010 PMID: 18782622

61. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, et al. Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol Evol. 2010; 2:425–40. doi: 10.1093/gbe/evq030 PMID: 20624745

62. Rota-Stabelli O, Telford MJ. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol Phylogenet Evol. 2008; 48(1):103–11. doi: 10.1016/j.ympev.2008.03.033 PMID: 18501642